Probability & Statistics Notes

B. Cordy

201-BLE, Winter 2021

Contents

Chapter 1

Foundations of Probability

Introduction: The Monty Hall Problem

In a certain popular game show, the contestant is offered a chance to win a car. She must choose one of three doors; behind one of the doors is a car, and behind the other two are goats. She has no evidence to help her determine which prize is behind which door, so assuming no collusion between the contestant and the game show producers, she has no reason to prefer any one door to any other.

After the contestant has made her choice, the host then opens a door. The host will never open the door that the contestant has chosen, and will always open a door with a goat behind it. This is possible since there are two doors with goats behind them. At this point, the host offers the contestant a chance to switch her choice from the door she initially chose to the other remaining unopened door.

How does the decision to switch doors affect her chance of winning the car?

- Switching *increases* her chance of winning from 1/3 to 1/2.
- Switching *increases* her chance of winning from 1/3 to 2/3.
- Switching *decreases* her chance of winning from 1/3 to 1/6.
- Her chance of winning is 1/3 regardless of whether she switches or not.
- Her chance of winning is 1/2 regardless of whether she switches or not.

This is a classic puzzle posed in a letter to the editors of American Statistician [7], later popularized in the 'Ask Marilyn' column of Parade magazine [8]. Debate over the correctness of the (correct) answer given there has raged ever since.

We have very strong intuitions about probability, which can sometimes be an aid and sometimes be a hindrance. In order to resolve the problem above in a manner which will be clear and convincing, even to individuals whose intuitions lead them to different conclusions, we need to establish some notation and fundamental laws we all agree on.

1.1 Sample Spaces & Events

Suppose that the outcome of some experiment is uncertain, but we can write down the set of all possible outcomes in advance. This set of possible outcomes is known as a *sample space*, and is the most basic object in the theory of probability. Precisely specifying a sample space will clarify the context of a problem and eliminate many potential sources of confusion.

Definition 1.1 *The set of possible outcomes of an experiment is known as a sample space, and typically denoted* Ω.

Example 1.1 *Consider a single roll of a single die. In this experiment, there are six possible outcomes, corresponding to the six distinct faces of the die. Thus, we take* $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Example 1.2 *Consider measuring the breaking strength (in Newtons) of a plank of wood. If we have no information about the size or shape of the plank, or about the precision with which the breaking strength will be measured, we should allow any positive value. Thus,* $\Omega = \{x \in \mathbb{R} \,|\, x \geq 0\}$.

Example 1.3 *Suppose that a coin is flipped, and then a marble is drawn from an urn which contains black, white, and grey marbles. In this scenario, we could take* $\Omega = \{HB, HW, HG, TB, TW, TG\}$*, using* HB *to represent flipping a head and drawing a black marble,* HW *to represent flipping a head and drawing a white marble, and so on.*

Example 1.4 *Suppose you play a game where a coin is flipped repeatedly until the first tail appears. We could take* $\Omega = \{T, HT, HHT, HHHT, HHHHT, \ldots\}$.

If the outcomes in a sample space can be listed (whether that list is finite or infinite) it is *discrete*, and if the set of all outcomes forms a continuum, as in the case of Example 1.2, then the sample space is *continuous*.

The choice of sample space depends on the experiment and on what question we as experimenters are trying to answer. If we roll a pair of dice, we might use the sample space $\Omega = \{2, 3, 4, \ldots, 11, 12\}$ to count the sum of the two resulting values, or $\Omega = \{(1,1), (1,2), (2,1), (2,2), \ldots, (5,6), (6,5), (6,6)\}$ to distinguish the dice and represent the values on each, so situations such as rolling doubles can be noted.

Another advantage of the larger sample space is that each of the outcomes is *equally likely*. Imagine one die is red and the other is blue (changing the colours of the dice certainly won't influence the results of the experiment). If neither die has any bias, then when the pair is rolled, each die is equally likely to land with any of its faces up. The red die showing 2 and blue showing 5 is as likely as the red die showing 6 and blue showing 6, and so on. Each ordered pair of outcomes (m, n) for $1 \leq m \leq 6$ and $1 \leq n \leq 6$ is as likely as any other.

If we write the number showing on the red die across the top row, the number showing on the blue die down the left column, and the sum of the two results in a table, we obtain the following.

$B \backslash R$	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Note that outcomes in the sample space $\Omega = \{2, 3, 4, \ldots, 11, 12\}$ are *not equally likely*. There are six ordered pairs $(1, 6)$, $(2, 5)$, $(3, 4)$, $(4, 3)$, $(5, 2)$, and $(6, 1)$ which yield a sum of seven, while only the single pair $(1, 1)$ gives a sum of two.

Events in a Sample Space

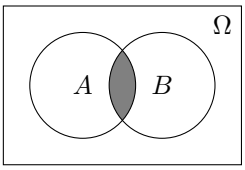
Definition 1.2 *Events are certain subsets of a sample space in which we might be interested, and are denoted by capital letters. We can describe an event using words or with more formal set notation.*

Example 1.5 *Suppose we roll a single die. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, let A be the event 'an even number is rolled' and let B be the event 'a number greater than four is rolled'. Then we can write $A = \{2, 4, 6\}$ and $B = \{5, 6\}$.*

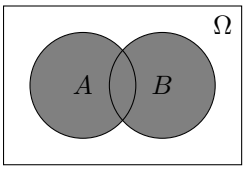
Example 1.6 *Consider a point selected at random inside a square with corners at $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$. Let $\Omega = \{(x, y) \,|\, 0 \leq x \leq 1$ and $0 \leq y \leq 1\}$, and let T be the event 'the point is in the top half of the square'. Then using set notation we would write $T = \{(x, y) \,|\, 0 \leq x \leq 1$ and $0.5 \leq y \leq 1\}$.*

Given two events $A, B \subseteq \Omega$, we can use the usual set operations of *intersection*, *union*, and *complement* to form the events $A \cap B$ (both A and B occur), $A \cup B$ (at least one of A and B occurs), and A^c (A does not occur). Another useful operation is the *set difference*, which is denoted $A \backslash B$, and defined by $A \backslash B = A \cap B^c$.

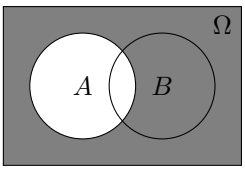
One can interpret $\cap$ as 'and', $\cup$ as 'or', and ^{c} as 'not', but remember that $\cup$ refers to the *inclusive or*, as in 'to participate you must be older than 16 or have permission of a parent' (*possibly both!*), not the *exclusive or*, as in 'the meal comes with a choice of soup or salad' (*not both!*).



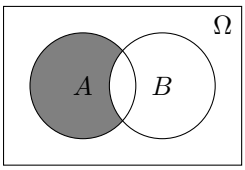
$A \cap B$



$A \cup B$



A^c



$A \backslash B$

The set operations described above satisfy many laws, which together make up the *algebra of sets*. Many of these laws are natural enough that there's likely no need for you to review them, for example $A \cup B = B \cup A$, or $(A^c)^c = A$. However, it's probably worth reviewing *DeMorgan's laws*.

$$(A \cap B)^c = A^c \cup B^c \quad \text{and} \quad (A \cup B)^c = A^c \cap B^c$$

We'll sometimes be dealing with long finite sequences of events $\{A_1, A_2, \ldots, A_n\}$ or infinite sequences $\{A_1, A_2, A_3, \ldots\}$ (also denoted $\{A_i\}_{i=1}^{\infty}$), so it's helpful to have some notation for abbreviating unions and intersections. We'll write large unions as below, and use similar notation for large intersections.

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \ldots \cup A_n \quad \text{and} \quad \bigcup_{i=1}^{\infty} A_i = A_1 \cup A_2 \cup A_3 \cup \ldots$$

Example 1.7 *In Example 1.4, a coin is flipped until the first tail appears. If A_k is the event 'exactly k heads are flipped', and E is the event 'a positive even number of heads are flipped', then $E = A_2 \cup A_4 \cup A_6 \cup \ldots = \bigcup_{i=1}^{\infty} A_{2i}$.*

Example 1.8 *Suppose that cards are drawn repeatedly from a deck until the first heart appears. Let H_i denote the event 'the i^{th} card drawn is a heart', and let T denote the event 'at least ten cards are drawn before the first heart appears'. Then we have $T = {H_1}^c \cap {H_2}^c \cap \ldots \cap {H_{10}}^c = \bigcap_{i=1}^{10} {H_i}^c$.*

Remark DeMorgan's laws apply to sequences as well. This means we could rewrite the event T in the example above as

$$T = \bigcap_{i=1}^{10} {H_i}^c = {H_1}^c \cap {H_2}^c \cap \ldots \cap {H_{10}}^c = (H_1 \cup H_2 \cup \ldots \cup H_{10})^c = \left(\bigcup_{i=1}^{10} H_i \right)^c.$$

Note $\bigcup_{i=1}^{10} H_i$ is the event 'at least one of the first ten cards drawn is a heart', so it's complement is 'none of the first ten cards drawn is a heart', which is T.

Let's return for a moment to rolling a pair of dice. Consider the events $N =$ 'the sum of the two values showing is nine', $F =$ 'one of the two dice shows a five', and $T =$ 'one of the two dice shows a two'. Suppose that we roll the dice and N occurs. Is it possible that F also occurred? Yes. If one die shows a five and the other shows a four, then both events have occurred. What about T? If N has occurred, it is

possible that T also occurred? This time the answer is no. If one of the dice shows two, the largest possible result would be eight. Thus, N and T cannot both occur when a pair of dice are rolled.

Definition 1.3 *Two events A and B in a sample space Ω are mutually exclusive if $A \cap B = \emptyset$, that is, if there are no outcomes in Ω where both events occur.*

Example 1.9 *Consider an experiment where a coin is flipped, and then a die is rolled, and let $\Omega = \{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$. The events H = 'a head is flipped' and E = 'an even number is rolled' are not mutually exclusive, since $H \cap E = \{H2, H4, H6\}$.*

Example 1.10 *Suppose you're waiting for a bus, and you measure the number of seconds it takes for the bus to arrive. Let E = 'the bus arrives in under 3 minutes' and F = 'the bus takes more than 5 minutes to arrive'. Then E and F are mutually exclusive. Formally, let $\Omega = \{x \in \mathbb{R} \,|\, x \geq 0\}$. In interval notation, we can write $E = [0, 180)$ and $F = (300, \infty)$, then clearly $E \cap F = \emptyset$.*

Extending the notion of mutually exclusive events given above in Definition 1.3, we call a finite or infinite sequence of events $\{A_1, A_2, A_3, ...\}$ mutually exclusive if for any pair of events A_i and A_j in the sequence, $A_i \cap A_j = \emptyset$.

Example 1.11 *In Example 1.4, a coin is flipped until the first tail appears. If A_k is the event 'exactly k heads are flipped', and B_k is the event 'at least k heads are flipped', then $\{A_1, A_2, A_3, ...\}$ is a sequence of mutually exclusive events, while $\{B_1, B_2, B_3, ...\}$ is not since, for example, $B_2 \cap B_3 = \{HHHT, HHHHT, ...\}$.*

1.2 Axioms of Probability

The saying goes 'you can't get something for nothing', and so it is in mathematics. However, you can get a lot for a little. The entirety of Euclidean plane geometry can be derived from a list of five simple statements. These foundational statements are known as axioms (greek for 'that which is evident'). In probability theory, we can make do with only three axioms, proposed by Kolmogorov [6] in the 1930s.

1. $P(A) \geq 0$ for any event A.
2. $P(\Omega) = 1$.
3. If $\{A_i\}_{i=1}^{\infty}$ is mutually exclusive, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

The first two axioms should be simple enough to understand at first glance, though the third, called the *axiom of countable additivity*, is not so clear. It states that in any mutually exclusive sequence of events, the probability that one (or more) of the events will occur can be obtained by adding up their probabilities.

Example 1.12 *Suppose we flip a coin until the first tail appears, as in Example 1.4, and consider the events* $A_1 = \{T\}$, $A_2 = \{HT\}$, $A_3 = \{HHT\}$, *and so on. In general,* $A_i =$ *'the first tail appears on the* i^{th} *flip'. Then the sequence* $\{A_i\}_{i=1}^{\infty}$ *is mutually exclusive since only one of the* A_i *can occur when the experiment is performed. In fact, the sequence* $\{A_i\}_{i=1}^{\infty}$ *is just the list of all possible outcomes,* Ω, *with each outcome considered a distinct event.*

Thus, $\bigcup_{i=1}^{\infty} A_i = \Omega$. *The third axiom then states that* $P(\Omega)$ *will be the result of adding up* $P(A_1) + P(A_2) + P(A_3) + \ldots$, *and by the second axiom,* $P(\Omega) = 1$, *so we can conclude, unsurprisingly, that the probabilities of all the individual outcomes in the sample space sum to one.*

Consequences of the Axioms

In this section, we'll derive some fundamental rules of probability. It's remarkable that the entire theory we'll develop rests ultimately on the three simple axioms which were given above.

Theorem 1.1 $P(\emptyset) = 0$.

Pf. Consider $\{E_i\}_{i=1}^{\infty}$ with $E_1 = \Omega$ and $E_k = \emptyset$ for $k > 1$. This sequence is mutually exclusive, since the intersection of any set with $\emptyset$ is $\emptyset$, so by the third axiom,

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

$$P(\Omega) = P(\Omega) + \sum_{i=2}^{\infty} P(\emptyset)$$

$$1 = 1 + \sum_{i=2}^{\infty} P(\emptyset)$$

$$0 = \sum_{i=2}^{\infty} P(\emptyset)$$

Since $P(\emptyset) \geq 0$ by the first axiom, we have a sum of nonnegative terms adding up to zero. Therefore, every term in the sum must be zero, i.e. $P(\emptyset) = 0$. □

Next, we would like to be able to use the third axiom on finite sequences of events as well as infinite sequences. To show this is possible, note that we can extend a finite sequence of events to an infinite one by tacking on an infinite sequence of empty events before appealing to the third axiom.

Theorem 1.2 *If* $\{A_i\}_{i=1}^{k}$ *is mutually exclusive,* $P(\bigcup_{i=1}^{k} A_i) = \sum_{i=1}^{k} P(A_i)$.

Pf. Consider the sequence $\{E_i\}_{i=1}^{\infty}$ with $E_i = A_i$ for $i \leq k$ and $E_k = \emptyset$ for $i > k$. This sequence is mutually exclusive since $\{A_i\}_{i=1}^{k}$ is a mutually exclusive sequence,

and the intersection of any set with $\emptyset$ is again $\emptyset$, so by the third axiom,

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

$$P\left(\bigcup_{i=1}^{k} A_i \cup \bigcup_{i=k+1}^{\infty} \emptyset\right) = \sum_{i=1}^{k} P(A_i) + \sum_{i=k+1}^{\infty} P(\emptyset)$$

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i) + \sum_{i=k+1}^{\infty} 0$$

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i)$$

□

Corollary 1.1 *$P(A^c) = 1 - P(A)$ for any event A.*

Pf. Observe that $\{A, A^c\}$ is a mutually exclusive sequence, and $A \cup A^c = \Omega$, so by Theorem 1.2, $P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$, and $P(\Omega) = 1$ by the second axiom. Thus, $P(A) + P(A^c) = 1$, i.e. $P(A^c) = 1 - P(A)$. □

Corollary 1.2 *If A and B are mutually exclusive, $P(A \cup B) = P(A) + P(B)$.*

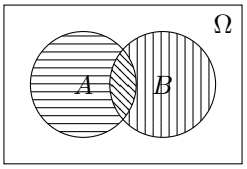
Pf. Apply Theorem 1.2 to the sequence $\{A_i\}_{i=1}^2$ with $A_1 = A$ and $A_2 = B$. □

Example 1.13 *Suppose that a card is drawn from a shuffled standard deck. What is the probability it is either a queen or not a face card?*

There are four queens in the deck, so $P(Q) = \frac{4}{52}$, and twelve face cards, so $P(F) = \frac{12}{52}$. By Corollary 1.1, $P(F^c) = 1 - \frac{12}{52} = \frac{40}{52}$, and since Q and F^c are mutually exclusive (no card can be both a queen and not a face card), by Corollary 1.2 we have $P(Q \cup F^c) = \frac{4}{52} + \frac{40}{52} = \frac{44}{52}$.

These two corollaries are very useful, but in practice, we'll need to be able to calculate $P(A \cup B)$ whether or not A and B are known to be mutually exclusive. To generalize the rule in Corollary 1.2, let's recall some useful properties of the set difference operation.

Two key points are that for any events A and B, (i) $A = (A \backslash B) \cup (A \cap B)$, and (ii) $A \cup B = (A \backslash B) \cup (B \backslash A) \cup (A \cap B)$. These identities are quite useful because the events on the right hand sides of both of these equations are *always* mutually exclusive. The case of (ii) is illustrated in the Venn diagram below, where $A \backslash B$ is shaded with horizontal lines, $B \backslash A$ with vertical lines, and $A \cap B$ with angled lines.



Theorem 1.3 *If $B \subseteq A$, then $P(B) \leq P(A)$.*

Pf. Consider the events $A\backslash B$ and $A \cap B$. As we noted above, these events are mutually exclusive, and their union is A. Thus, $P(A) = P(A\backslash B) + P(A \cap B)$ by Corollary 1.2, and if $B \subseteq A$, then $A \cap B = B$, so $P(A) = P(A\backslash B) + P(B)$. This implies $P(B) \leq P(A)$ since $P(A\backslash B) \geq 0$ by the first axiom. □

The last fact we'll derive is the extremely useful *inclusion-exclusion principle*, which allows us to compute the probability $A \cup B$ will occur without making any assumptions about the events A and B.

Theorem 1.4 *(Inclusion-Exclusion)* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Pf. Note $A = A\backslash B \cup (A \cap B)$, and since $A\backslash B$ and $A \cap B$ are mutually exclusive, $P(A) = P(A\backslash B) + P(A \cap B)$ by Corollary 1.2. After isolating $P(A\backslash B)$, we obtain $P(A\backslash B) = P(A) - P(A \cap B)$. Similarly, $P(B\backslash A) = P(B) - P(A \cap B)$.

$$\begin{aligned} P(A \cup B) &= P((A\backslash B) \cup (B\backslash A) \cup (A \cap B)) \\ &= P(A\backslash B) + P(B\backslash A) + P(A \cap B) \\ &= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

□

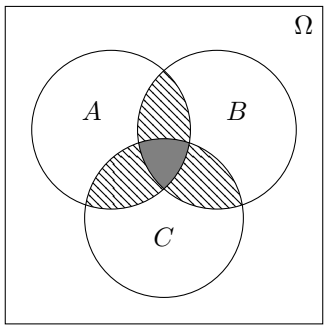
Though the argument above is fairly opaque, the inclusion-exclusion principle itself is quite intuitive. To illustrate, suppose we draw a card from a standard deck, let H = 'the card is a heart' and F = 'the card is a face card'. Then to determine the probability that the card drawn is either a heart or a face card (that is, $H \cup F$ occurs), we should count up all of the hearts, of which there are 13, and count up all the face cards, of which there are 12.

However, the three cards in $H \cap F = \{J\heartsuit, Q\heartsuit, K\heartsuit\}$ were counted twice. If we correct for this, the number of cards in $H \cup F$ is $13 + 12 - 3 = 22$, and therefore we have $P(H \cup F) = \frac{22}{52} = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = P(H) + P(F) - P(H \cap F)$.

Example 1.14 *Suppose that we toss a pair of dice. What is the probability that a two appears on the face of a die, or the sum of the values on both dice is 6?*

Let A = 'a two appears on a die' and B = 'the total rolled is six'. From the table constructed in the previous section, we can tell that $P(A) = \frac{11}{36}$, and $P(B) = \frac{5}{36}$. Furthermore, $A \cap B = \{(2,4),(4,2)\}$, so $P(A \cap B) = \frac{2}{36}$. Thus, by the inclusion-exclusion principle, $P(A \cup B) = \frac{11}{36} + \frac{5}{36} - \frac{2}{36} = \frac{14}{36}$.

The inclusion-exclusion principle can be extended to unions of three events or more. In the case of three, if we write $P(A \cup B \cup C) = P(A) + P(B) + P(C)$ then we have double counted the outcomes in the lined areas in the Venn diagram below, and triple counted the outcomes in $A \cap B \cap C$ (the shaded central area).



To correct for double counting the lined areas, we can subtract the probabilities of $A \cap B$, $B \cap C$, and $C \cap A$, but each of these includes $A \cap B \cap C$. It was counted three times and has now been subtracted out three times. Thus, we need to add it back in. All together we obtain the formula below.

$$\begin{aligned} P(A \cup B \cup C) = & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(B \cap C) - P(C \cap A) \\ & + P(A \cap B \cap C) \end{aligned}$$

This pattern continues. In the case of four events A, B, C, and D, to determine $P(A \cup B \cup C \cup D)$, we would add up the probabilities of each event, subtract the double intersections, add back in the triple intersections, and finally subtract out the quadruple intersection $A \cap B \cap C \cap D$.

The Basic Probability Laws

Three of the results we've derived above are such fundamental ingredients in most of the probability calculations we'll do that it's worth having them written down in one place. Let's call these the *Basic Probability Laws*.

- $\boxed{P(A \cup B) = P(A) + P(B) \text{ if } A \text{ and } B \text{ are mutually exclusive.}}$
- $\boxed{P(A^c) = 1 - P(A) \text{ for any event } A.}$
- $\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B) \text{ for any events } A \text{ and } B.}$

1.3 What is a Probability?

The Kolmogorov axioms and the algebra of sets provide us with a mathematical model we can use to reason about the probability of various events in the world occurring. However, we never explicitly defined what is meant by stating that the probability of some event is equal to some number. What does this number really

mean? In fact, there are many points of view on this philisophical question, three of the most common are described below.

The Classical Interpretation: In an experiment having n equally likely outcomes, stating that $P(A) = \frac{k}{n}$ means simply that A occurs in k of these outcomes.

This is a concise and simple definition, but using it means assuming in advance, without any empirical evidence, that the experiment we're performing has n equally likely outcomes. In some cases this is unrealistic (many dice have dimples drilled in them, making the lighter faces with higher numbers more likely), and in others it simply doesn't apply at all (an election does not have n equally likely outcomes).

The Frequentist Interpretation: If we perform exactly the same experiment n times, and R_n is the ratio of the number of times A occurred to the total number of trials, then $P(A) = \lim_{n\to\infty} R_n$.

This interpretation defines probability as long term behaviour in a repeatable experiment. We make no assumptions about the likelihood of any outcome. The gripe with this definition is that some experiments cannot be performed repeatedly under the same conditions (an election or a sports match for example), and even in an experiment that can be, it's not possible to actually perform infinitely many trials and determine $\lim_{n\to\infty} R_n$. The interpretation is conceptually nice, but not immediately helpful for calculations.

The Bayesian Interpretation: Probability represents a degree of confidence in the correctness of a proposition, subject to a set of basic logical laws and to the current state of knowledge of the individual assigning the probabilities.

This is often referred to as the subjectivist interpretation, since two individuals with different states of knowledge can calculate different probabilities of the same event occurring. This may sound strange, but if one individual knows that a certain die is loaded so that rolling a six is less likely, and another doesn't, then they *should* calculate different probabilities of a six being rolled.

As it turns out, this dependence on the state of knowledge of an agent can be a very desirable feature in practice, but it makes the Bayesian interpretation more conceptually complicated and subtle than the other two above.

Throughout the rest of this chapter, the frequentist interpretation will give you a simple conceptual basis for thinking about probability, so I encourage you to have that interpretation in the back of your mind most of the time. Note that proponents of any of the three interpretations would agree on the correctness of the Kolmogorov axioms, and hence on the entire theory we'll develop in this chapter.

Continuity & Events with Probability Zero

A finite or infinite sequence of events is called *increasing* when $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$, and *decreasing* when $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$. Informally, each step along an increasing sequence adds zero or more outcomes, and each step along a decreasing sequence removes zero or more outcomes.

Example 1.15 *Suppose we flip a coin five times, and define A_i = 'the first i flips are all heads'. Then $\{A_i\}_{i=1}^5$ is a decreasing sequence since A_1 includes all outcomes where the first flip is a head, A_2 includes only outcomes where both of the first two flips are heads, and so on, with fewer outcomes at each step, until A_5 includes only the single outcome where every flip is a head.*

Example 1.16 *Consider a dart being thrown at a dartboard, and let D_i denote the event 'the dart lands within i millimeters of the center', then the sequence $\{D_i\}_{i=1}^\infty$ is an increasing sequence of events.*

Theorem 1.5 *(Continuity of the Probability Function) If $\{A_i\}_{i=1}^\infty$ is an increasing sequence of events, then*

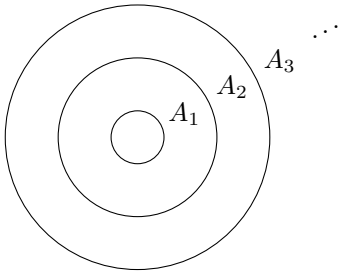
$$P(\bigcup_{i=1}^{\infty} A_i) = \lim_{i\to\infty} P(A_i).$$

Similarly, if $\{A_i\}_{i=1}^\infty$ is an decreasing sequence of events, then

$$P(\bigcap_{i=1}^{\infty} A_i) = \lim_{i\to\infty} P(A_i).$$

Pf. Assume $\{A_i\}_{i=1}^\infty$ is an increasing sequence of events (the decreasing case is left as an exercise). We'll use the same trick we used to prove the inclusion-exclusion principle; we'll create a new sequence of events $\{E_i\}_{i=1}^\infty$ which is mutually exclusive, but for which $\cup_{i=1}^\infty A_i = \cup_{i=1}^\infty E_i$.

Let $E_1 = A_1$, $E_2 = A_2 \backslash A_1$, $E_3 = A_3 \backslash A_2$, and so on. This way, if an outcome appears in the sequence $\{A_i\}_{i=1}^\infty$ for the *first time* in A_k, then it appears *only* in E_k and in no other event in the sequence $\{E_i\}_{i=1}^\infty$.



To illustrate, you can visualize $\{A_i\}_{i=1}^\infty$ in a Venn diagram as a sequence of disks, with each disk surrounding and *including* those that came before. On the

other hand, $\{E_i\}_{i=1}^{\infty}$ is the sequence of *rings*, which don't overlap, so this sequence is mutually exclusive. Every outcome in A_n is also in E_n, or in some E_i with $i < n$, so $\cup_{i=1}^{n} A_i = \cup_{i=1}^{n} E_i$ for all n (and $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} E_i$ for the same reason).

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(E_i) = \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} E_i\right)$$

Note that the second equality is justified by axiom (iii), the third is the definition of an infinite sum, and the fourth is axiom (iii) being used again, this time in the opposite direction. To finish the argument, recall that $\cup_{i=1}^{n} A_i = \cup_{i=1}^{n} E_i$ and $\{A_i\}_{i=1}^{\infty}$ is an increasing sequence, so $\cup_{i=1}^{n} A_i = A_n$.

$$\lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} E_i\right) = \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} A_i\right) = \lim_{n\to\infty} P(A_n)$$

□

Example 1.17 *Suppose we pick a real number in the* $[0, 1]$ *by choosing each of its decimal digits at random. What is the probability the number selected is* $\frac{2}{9}$ *?*

If we let T_i *be the event 'the first* i *digits are all twos', then the sequence* $\{T_i\}_{i=1}^{\infty}$ *is a decreasing sequence, and the event* $\cap_{i=1}^{\infty} T_i$ *occurs when all the* T_i*'s occur, i.e. when the number selected is* $0.2222222... = \frac{2}{9}$ *.*

$$P(T_1) = \frac{1}{10}, \quad P(T_2) = \frac{1}{100} = \frac{1}{10^2}, \quad P(T_3) = \frac{1}{1000} = \frac{1}{10^3}, \quad ...$$

In general, $P(T_n) = \frac{1}{10^n}$*, so by Theorem 1.5 above (for decreasing sequences), the probability that the chosen number is* $\frac{2}{9}$ *is*

$$P(\bigcap_{i=1}^{\infty} T_i) = \lim_{n\to\infty} P(T_n) = \lim_{n\to\infty} \frac{1}{10^n} = 0.$$

This example illustrates the difference between an event with probability zero and an event which can never occur. If we choose a number in the interval $[0, 1]$ at random, *every possible outcome can be shown to have probability zero.*

To get some intuition for why this is the case, suppose I've chosen a number, and you decide to confirm whether or not it's equal to $\frac{2}{9}$. You ask 'does the first digit match', and I reply 'yes'; you ask 'does the second digit match', and I reply 'yes'; and so on. If I have truly chosen each digit at random, then every time you ask about the next digit, I have a $\frac{9}{10}$ chance of replying 'no', which will end the game immediately, and whenever I reply 'yes' you simply challenge me again.

Note that if we let $N_{\frac{2}{9}}$ denote the event 'the number selected is not $\frac{2}{9}$', then by Corollary 1.1, $P(N_{\frac{2}{9}}) = 1$. In fact, for any $x \in [0, 1]$ if we define N_x in the same way, then $P(N_x) = 1$ for all $x \in [0, 1]$. Each event N_x has probability 1, but it is certain that one of them will *not* occur. We say that events with probability one *almost surely* occur and events with probability zero *almost never* occur.

Remark The subtle distinction between an event with probability zero and an event which never occurs disappears in a finite sample space. In that case, an event with probability zero never occurs, and an event with probability one always occurs.

In any *finite* number of trials of an experiment, you should expect an almost sure event to always occur, and an event which almost never occurs to actually never occur. This is the frequentist interpretation of events with probability one and probability zero. You should be willing to take a bet that an almost sure event will occur with *any odds*, no matter how large the required wager and how small the potential gain. Conversely, you should never be willing to take a bet on an event which almost never occurs, no matter how small the required wager and how large the potential gain.

1.4 Conditional Probability

Until now, our probability calculations have been done with no information other than a precise description of all possible outcomes of an experiment, given in the form of a sample space, and the assumption that certain events are equally likely, such as each of the six outcomes that can occur when a die is rolled.

In practice, we are often given incomplete information about the result of an experiment, and we need to be able to update the probability that an event occurs when presented with this information. Consider the following examples.

Example 1.18 *Suppose a single die is rolled. We are not told the result, but we are told it was an even number. What is the probability the result was six?*

Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, let $S =$ 'the outcome was a six', and let $E =$ 'the outcome was an even number'. Initially, each outcome in Ω is equally likely, and $P(S) = \frac{1}{6}$. However, given the knowledge that an even number was rolled, the sample space has been reduced from all the outcomes in Ω to only those in E. Now each of the three even outcomes are equally likely, and each of the odd outcomes has probability zero. We write $P(S \mid E) = \frac{1}{3}$ to indicate the probability of S under the assumption that E has occurred is $\frac{1}{3}$.

Example 1.19 *In a certain school, 35% of students receive an A grade in reading, 22% of students receive an A grade in mathematics, and 16% of students receive an A grade in both. Of those students who received an A grade in reading, what proportion also received an A grade in mathematics?*

Let the number of students at the school be denoted by n. Then the number that received an A in reading is $0.35n$, the number that received an A in mathematics is $0.22n$, and the number that received an A in both is $0.16n$. If we are considering only students who received an A in reading, we are dealing with a pool of $0.35n$ students. The only students in this pool who received an A in math must have received an A in both subjects, so there are total of $0.16n$ such students.

Thus, the proportion of A students in reading who also received an A in math is $\frac{0.16n}{0.35n} = \frac{0.16}{0.35} \simeq 0.46 = 46\%$. *Note that if we select a student at random and define the events* $R =$ *'the student received an A in reading' and* $M =$ *'the student received an A in math', then we can rewrite this calculation as*

$$P(M \mid R) = \frac{P(M \cap R)}{P(R)} = \frac{0.16}{0.35} \simeq 0.46.$$

We'll use this idea to formally define conditional probability, but it's often more productive to view conditional probability as a reduction of the sample space. In the example above, if you gathered *only students who received an A in reading* in a room together and asked those who received an A in math to raise their hands, 46% of them would raise their hands (if they're honest).

Definition 1.4 *Given two events* E *and* F *in some sample space* Ω*, the conditional probability of* E *given* F *is denoted* $P(E \mid F)$*, and defined by*

$$\boxed{P(E \mid F) = \frac{P(E \cap F)}{P(F)}.}$$

Note that division by zero could be problematic here, so we'll only define the conditional probability of E given F when $P(F) \neq 0$.

The Intersection Law

This definition can be used either to compute $P(E \mid F)$ when $P(E \cap F)$ and $P(F)$ are known, or to compute $P(E \cap F)$ when $P(E \mid F)$ and $P(F)$ are known. In fact, we'll often use it in this second way, so it's worth renaming some variables and moving terms around to make it easier to use in that direction.

$$\begin{aligned} P(B \mid A) &= \frac{P(B \cap A)}{P(A)} \\ P(B \mid A)P(A) &= P(B \cap A) \\ P(B \mid A)P(A) &= P(A \cap B) \\ \boxed{P(A \cap B) &= P(A)P(B \mid A)} \end{aligned}$$

Intuitively, if A and B both occur, then A must occur and B must occur in this new setting where the event A has already happened.

Example 1.20 *Suppose we draw two cards from a shuffled deck without replacement. What is the probability that both are hearts?*

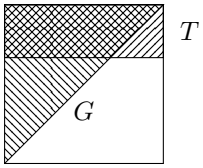
Let $H_1 =$ *'the first card drawn is a heart' and* $H_2 =$ *'the second card drawn is a heart'. Then both cards are hearts when* $H_1 \cap H_2$ *occurs.*

$$P(H_1 \cap H_2) = P(H_1)P(H_2 \mid H_1) = \frac{13}{52} \cdot \frac{12}{51} = \frac{156}{2652} \simeq 0.059$$

The key point is that if H_1 has occurred, a heart has been removed from the deck. There are then 51 cards remaining, 12 of which are hearts, so $P(H_2 \mid H_1) = \frac{12}{51}$.

Example 1.21 *Consider a point selected at random in a unit square, as in Example 1.6. We take $\Omega = \{(x, y) \mid 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1\}$. If the point has $x < y$, what is the probability it lies in the top third of the square?*

We're looking for $P(T \mid G)$ where $T = \{(x, y) \mid 0 \leq x \leq 1 \text{ and } \frac{2}{3} \leq y \leq 1\}$ and $G = \{(x, y) \mid 0 \leq x < y \leq 1\}$. To compute this probability, we can draw a picture.



The square has unit area, and hence the area of any region inside it represents the probability a randomly selected point will lie in that region.

$$P(T \cap G) = \frac{1}{3} - \frac{1}{18} = \frac{5}{18} \quad \textit{and} \quad P(G) = \frac{1}{2}$$

Therefore, $P(T \mid G) = \dfrac{P(T \cap G)}{P(G)} = \dfrac{\frac{5}{18}}{\frac{1}{2}} = \dfrac{5}{9}$.

As an application of these ideas, let's analyze a classic puzzle presented by Martin Gardner [3]. Consider the two statements below.

1. Mr. Jones has two children. The older child is a girl. What is the probability his other child is a girl?
2. Mr. Smith has two children. At least one is a girl. What is the probability the other child is a girl?

Answering the first question is relatively straightforward. Consider the sample space $\Omega = \{bb, bg, gb, gg\}$, the first letter representing the gender of the elder child, and the second letter representing the gender of the younger. Given no information about Mr. Jones' family, each outcome in this sample space is equally likely. If we let $E =$ 'the elder child is a girl' and $Y =$ 'the younger child is a girl', then

$$P(Y \mid E) = \frac{P(Y \cap E)}{P(E)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}.$$

Which agrees with intuition. The second question involving Mr. Smith's family is more problematic. On the one hand, we could let $A =$ 'at least one child is a girl' and $B =$ 'both children are girls', and calculate

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

On the other hand, without any information, a given child is equally likely to be a boy or girl, and the fact that one child is a girl gives us no information about Mr. Smith's *other child*, so the chance that second child is a girl should still be $\frac{1}{2}$.

Which is correct, $\frac{1}{3}$ or $\frac{1}{2}$? The key to resolving the paradox is to clarify what experiment is actually being performed. If we *choose at random a family* with two children, one of which is a girl, the three possibilities bg, gb, and gg are equally likely, and our calculation of $P(B \mid A)$ above is correct.

However, if we *choose at random a girl* from a family with two children, a bg or gb family is half as likely to be chosen as a gg family, since the latter has twice as many girls! Thus, $P(\{bg\}) = P(\{gb\}) = \frac{1}{4}$, and $P(\{gg\}) = \frac{1}{2}$. In this context, the event $A =$ 'at least one child is a girl' always occurs, and we can calculate

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{\frac{1}{2}}{1} = \frac{1}{2}.$$

The moral of the story is that we need to specify the experiment being performed and the source of uncertainty; in this case, the process by which Mr. Smith's family was chosen. When you read 'Mr. Smith has two children', you should immediately ask who Mr. Smith is and how he was chosen. Without this information, the second question is ambiguous.

Extending the Intersection Law

We've derived the intersection law $P(A \cap B) = P(A)P(B \mid A)$ from the definition of conditional probability. It's worth noting that this law generalizes to intersections of any number of events,

$$P(A_1 \cap A_2 \cap A_3 \cap ...) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2)...$$

where each factor on the right is the conditional probability of A_i calculated under the assumption that all prior events in the sequence have occurred.

Example 1.22 *What is the probability that when we draw four cards from a deck without replacement, all four cards are different suits?*

Let $D_i =$ 'the i^{th} card is a different suit from all those cards the came before', then $P(D_1) = 1$, since no cards were drawn before the first, and we calculate

$$\begin{aligned} P(D_1 \cap D_2 \cap D_3 \cap D_4) &= P(D_1)P(D_2 \mid D_1)P(D_3 \mid D_1 \cap D_2)P(D_4 \mid D_1 \cap D_2 \cap D_3) \\ &= 1 \cdot \frac{39}{51} \cdot \frac{26}{50} \cdot \frac{13}{49} \simeq 0.1055 \end{aligned}$$

Example 1.23 *Suppose that seven lightbulbs are in a box, but only two work. Bob will randomly select lightbulbs one at a time and test them. What is the probability he'll find both working lightbulbs after exactly three tests?*

Let W_i = 'the i^{th} lightbulb tested works'. Then we need W_3 to occur, and one of W_1 and W_2. If W_1 and W_2 both occur then Bob will be finished after only two tests.

$$\begin{aligned}
&P((W_1 \cap {W_2}^c \cap W_3) \cup ({W_1}^c \cap W_2 \cap W_3)) \\
&= P(W_1 \cap {W_2}^c \cap W_3) + P({W_1}^c \cap W_2 \cap W_3) \\
&= P(W_1)P({W_2}^c \,|\, W_1)P(W_3 \,|\, W_1 \cap {W_2}^c) + P({W_1}^c)P(W_2 \,|\, {W_1}^c)P(W_3 \,|\, {W_1}^c \cap W_2) \\
&= \frac{2}{7} \cdot \frac{5}{6} \cdot \frac{1}{5} + \frac{5}{7} \cdot \frac{2}{6} \cdot \frac{1}{5} = \frac{10}{210} + \frac{10}{210} \simeq 0.095
\end{aligned}$$

Note that $W_1 \cap {W_2}^c \cap W_3$ and ${W_1}^c \cap W_2 \cap W_3$ are mutually exclusive, so we can pass from the first to the second line without using the inclusion-exclusion principle.

Conditional Probability is Well-Behaved

Interpreting conditional probability as a reduction of the sample space from Ω to some $B \subseteq \Omega$ with $P(B) > 0$, it's not hard to convince yourself that the Kolmogorov axioms hold for events conditioned on the occurrence of B. The meaning of each axiom is the same, it's just being interpreted in the smaller sample space B.

1. $P(A \,|\, B) \geq 0$.
2. $P(B \,|\, B) = 1$.
3. If $\{A_i\}_{i=1}^{\infty}$ is mutually exclusive, $P\left(\bigcup_{i=1}^{\infty} A_i \,\middle|\, B\right) = \sum_{i=1}^{\infty} P(A_i \,|\, B)$.

This implies that *all the laws we derived in Section 1.2 apply for conditional probabilities as well*. The inclusion-exclusion law, for example, can be written for events A and B conditioned on C as

$$P(A \cup B \,|\, C) = P(A \,|\, C) + P(B \,|\, C) - P(A \cap B \,|\, C).$$

Example 1.24 *Suppose that in a certain area, 36% of all cats are black, and 76% of all black cats are mean. What is the probability that a randomly selected cat is black and not mean?*

$$P(B \cap M^c) = P(B)P(M^c \,|\, B) = P(B)(1 - P(M \,|\, B)) = 0.36 \cdot 0.24 = 0.0864$$

Remark If you read up on conditional probability, you might encounter notation such as $P(A \,|\, B, C)$. This means we are assuming all events on the right side of the vertical line have occurred, so $P(A \,|\, B, C) = P(A \,|\, B \cap C)$.

1.5 The Law of Total Probability

If two cards are drawn from a shuffled deck, what is the probability the second card drawn is a heart? At first, it's tempting to reply that in order to answer, we would require more information, namely, whether the first card drawn was a heart or not.

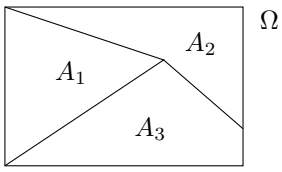
However, recalling the frequentist interpretation, the question above is perfectly well posed. If we repeat the experiment a very large number of times, then in some proportion of trials, the second card will be heart. In order to compute this value analytically, we'll need the result below.

Theorem 1.6 *(Law of Total Probability) If $\{A_i\}_{i=1}^n$ is any mutually exclusive sequence of events with $\bigcup_{i=1}^n A_i = \Omega$, then for any event B,*

$$\boxed{P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \ ... \ + P(B \mid A_n)P(A_n).}$$

Note that we require $P(A_i) > 0$ for each A_i, so that the conditional probabilities on the right are defined.

When a sequence $\{A_i\}_{i=1}^n$ is mutually exclusive and has $\bigcup_{i=1}^n A_i = \Omega$, we say that the sequence *partitions the sample space*. This means that every possible outcome is in exactly one of the A_i.



Corollary 1.3 *If A is an event with $0 < P(A) < 1$, then for any event B,*

$$\boxed{P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c).}$$

Pf. Apply Theorem 1.6 to the sequence $\{A, A^c\}$, which always partitions Ω. □

Example 1.25 *If two cards are drawn from a shuffled deck, what is the probability the second card drawn is a heart?*

Let H_1 = 'the first card is a heart' and H_2 = 'the second card is a heart'. We'll apply the corollary above to compute the probability of H_2.

$$\begin{aligned} P(H_2) &= P(H_2 \mid H_1)P(H_1) + P(H_2 \mid {H_1}^c)P({H_1}^c) \\ &= \frac{12}{51} \cdot \frac{1}{4} + \frac{13}{51} \cdot \frac{3}{4} = \frac{51}{204} = \frac{1}{4} \end{aligned}$$

This result makes intuitive sense. We're computing $P(H_2)$ in the absence of any information about the first card, so we may as well have drawn that card and put it on the bottom of the deck without looking at it.

Pf. (of Theorem 1.6) If the sequence $\{A_i\}_{i=1}^n$ partitions Ω, and $B \subseteq \Omega$,

$$\begin{aligned} \Omega &= A_1 \cup A_2 \cup \ ... \ \cup A_n \\ B \cap \Omega &= B \cap (A_1 \cup A_2 \cup \ ... \ \cup A_n) \\ B &= (B \cap A_1) \cup (B \cap A_2) \cup \ ... \ \cup (B \cap A_n). \end{aligned}$$

The sequence of events $B \cap A_1, B \cap A_2, \ldots, B \cap A_n$ is mutually exclusive because $\{A_i\}_{i=1}^n$ was mutually exclusive, and this new sequence is obtained by reducing the number of outcomes in each event. The events had empty intersections initially, so removing outcomes will keep the intersections empty.

$$\begin{aligned} P(B) &= P((B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n)) \\ &= P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_n) \\ &= P(A_1 \cap B) + P(A_2 \cap B) + \ldots + P(A_n \cap B) \\ &= P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \ldots + P(A_n)P(B \mid A_n) \\ &= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \ldots + P(B \mid A_n)P(A_n) \end{aligned}$$

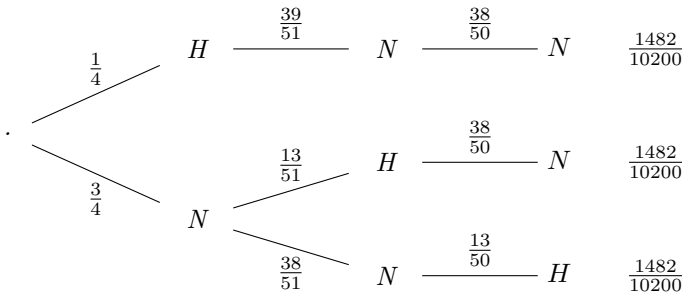
□

Tree Diagrams

Tree diagrams are a nice graphical tool for organizing calculations, and have the law of total probability built in to their structure. Each column represents a step in an experiment. If we draw three balls from an urn, for example, each column will represent a draw. If we roll a die and then flip a coin, the first column will represent the roll and the second will represent the flip.

Example 1.26 *What is the probability that, when three cards are drawn from a shuffled deck without replacement, exactly one heart appears?*

In the tree diagram below, the three columns represent three draws from a deck of cards, when a draw results in a heart, we denote that with an H, and when a draw does not result in a heart, we denote that with an N.



Each edge is labelled with a conditional probability. For example, the top-center edge which runs from H to N is labelled with $P(N_2 \mid H_1)$, the probability a non-heart is drawn on the second draw, after drawing a heart on the first. Note that we only depict the branches where the desired event ('at least one heart appears') occurs. To find the probability of this event, we simply multiply along each branch, and add the

results. If $E =$ 'exactly one heart is drawn', then

$$P(E) = \frac{1482}{10200} + \frac{1482}{10200} + \frac{1482}{10200} = \frac{4446}{10200} \simeq 43.6\%.$$

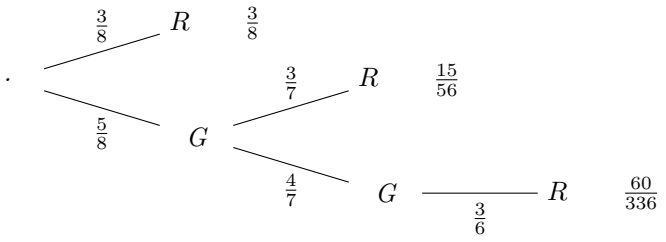
This approach works because the branches represent mutually exclusive events (at each fork, a given trial of the experiment can proceed along only one path), so we can add the probabilities of each branch. Furthermore, the intersection law states that we can compute the probability of all events along a branch occurring by multiplying their conditional probabilities as we did above. Note that if we applied the law of total probability in the above example, partitioning the sample space with $\{H_1 \cap H_2, H_1 \cap N_2, N_1 \cap H_2, N_1 \cap N_2\}$, we would obtain

$$\begin{aligned} P(E) &= P(E \,|\, H_1 \cap H_2)P(H_1 \cap H_2) + P(E \,|\, H_1 \cap N_2)P(H_1 \cap N_2) \\ &\quad + P(E \,|\, N_1 \cap H_2)P(N_1 \cap H_2) + P(E \,|\, N_1 \cap N_2)P(N_1 \cap N_2) \\ &= 0 + P(E \,|\, H_1 \cap N_2)P(N_2 \,|\, H_1)P(H_1) + P(E \,|\, N_1 \cap H_2)P(H_2 \,|\, N_1)P(N_1) \\ &\quad + P(E \,|\, N_1 \cap N_2)P(N_2 \,|\, N_1)P(N_1) \\ &= P(N_3 \,|\, H_1 \cap N_2)P(N_2 \,|\, H_1)P(H_1) + P(N_3 \,|\, N_1 \cap H_2)P(H_2 \,|\, N_1)P(N_1) \\ &\quad + P(H_3 \,|\, N_1 \cap N_2)P(N_2 \,|\, N_1)P(N_1) \end{aligned}$$

and this is exactly how we calculated the result above. Each path from left to right corresponds to a term in the sum above. The event E never occurs if two hearts are drawn, so $P(E \,|\, H_1 \cap H_2) = 0$ and that term (which corresponds to a branch not depicted in the diagram) vanishes. Although working through the notation is good practice, I expect you'll agree that in cases like this, the tree diagram is more intuitive and simplifies the bookkeeping.

If at any point in the tree diagram, the event we're interested in is sure to occur, there's no need to continue along that branch any further. We can save time in this way, as illustrated in the example below.

Example 1.27 *Suppose that an urn contains three red and five green balls. If three balls are drawn from the urn without replacement, what is the probability that at least one red ball will be drawn?*

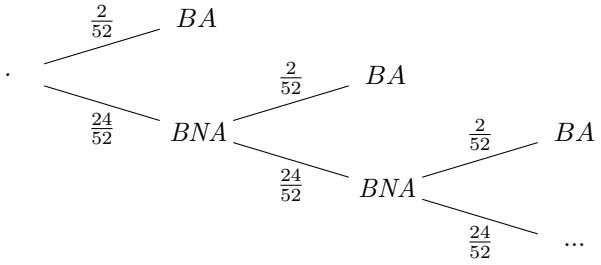


As soon as a red ball is drawn, there's no need to continue the branch. Regardless of what happens from that point on, a red ball will have been drawn. If $A =$ 'at least

one red ball is drawn', then

$$P(A) = \frac{3}{8} + \frac{15}{56} + \frac{60}{336} = \frac{276}{336} \simeq 82.1\%.$$

Example 1.28 *Suppose we draw cards with replacement from a shuffled deck until the first ace appears. What is the probability that all cards drawn are black?*



Here BA denotes 'black ace' and BNA denotes 'black non-ace'. There's no limit to the number of draws that could occur before the first black ace appears, so the tree diagram is infinite. However, we can still sum the products along each branch. Let OB = 'only black cards are drawn', then

$$\begin{aligned} P(OB) &= \frac{2}{52} + \frac{24}{52}\cdot\frac{2}{52} + \frac{24}{52}\cdot\frac{24}{52}\cdot\frac{2}{52} + \frac{24}{52}\cdot\frac{24}{52}\cdot\frac{24}{52}\cdot\frac{2}{52} + \ ... \\ &= \frac{2}{52}\left(1 + \frac{24}{52} + \left(\frac{24}{52}\right)^2 + \left(\frac{24}{52}\right)^3 + \ ...\right) \\ &= \frac{2}{52}\left(\sum_{i=0}^{\infty}\left(\frac{24}{52}\right)^n\right) \\ &= \frac{2}{52}\cdot\frac{1}{1-(\frac{24}{52})} = \frac{1}{14} \simeq 7.1\% \end{aligned}$$

Note that the sum of the infinite series above was evaluated with the geometric series sum formula, $a + ar + ar^2 + ... = \frac{a}{1-r}$ when $-1 < r < 1$.

1.6 Independence

Suppose we flip a coin and then roll a die. Let H denote the event 'a head is flipped', E denote the event 'and even number is rolled' and L denote the event 'a number larger than three was rolled'. Note that

$$P(E) = \frac{1}{2} \ \text{ and } \ P(E \,|\, H) = \frac{1}{2}.$$

Knowledge that a head was flipped doesn't change the probability an even number was rolled. On the other hand, we can calculate

$$P(E) = \frac{1}{2} \ \text{ and } \ P(E \,|\, L) = \frac{2}{3}.$$

In this case, knowledge that the number rolled was a four, five, or six increases the probability that the result of the roll was even. The events E and H are said to be *independent*, while the events E and L are *not independent*.

Definition 1.5 *Events A and B are called independent if $P(B) = P(B \mid A)$.*

Example 1.29 *Suppose we flip a coin two times. Note that $\Omega = \{HH, HT, TH, TT\}$ is a sample space with equally likely outcomes for the experiment.*

If we consider the events H_1 = 'the first flip is a head' and H_2 = 'the second flip is a head', then $P(H_2) = \frac{1}{2}$ and $P(H_2 \mid H_1) = \frac{1}{2}$, so these events are independent.

If we instead let H_1 = 'the first flip is a head' and T_1 = 'the first flip is a tail', then $P(T_1) = \frac{2}{4}$ and $P(T_1 \mid H_1) = 0$, so these two events are not independent.

Example 1.30 *Consider two cards drawn from a shuffled deck. Let S_1 = 'a spade is drawn on the first draw' and S_2 = 'a spade is drawn on the second draw'. Then S_1 and S_2 are independent provided the draws are done with replacement, but not independent if the draws are done without replacement. In this second case, $P(S_2 \mid S_1) = \frac{12}{51}$, but using the law of total probability, $P(S_2) = \frac{1}{4}$.*

Using the definition of the conditional probability $P(A \mid B)$, we can state what it means for two events to be independent in a different way.

$$P(A) = P(A \mid B) \;\leftrightarrow\; P(A) = \frac{P(A \cap B)}{P(B)} \;\leftrightarrow\; P(A)P(B) = P(A \cap B)$$

Thus, when A and B are independent events, $P(A \cap B) = P(A)P(B)$. This relation is often used as the definition of independence. As we just saw, it's equivalent to the definition we gave earlier but also applies when $P(A) = 0$ or $P(B) = 0$ (in these cases $P(A \mid B)$ or $P(B \mid A)$ would be undefined).

Remark Independence is a symmetric relation, which you can interpret intuitively as saying 'the probability of one event occurring is not influenced by the occurrence of the other' without worrying about which event comes first. This revised definition confirms that intuition, as swapping the roles of A and B yields the same equation.

Theorem 1.7 *If A and B are independent, then A and B^c are independent.*

Pf. Note that $A = (A \cap B) \cup (A \cap B^c)$ and hence $P(A) = P(A \cap B) + P(A \cap B^c)$ since the two events on the right are mutually exclusive. Isolating $P(A \cap B^c)$,

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) \\ &= P(A)(1 - P(B)) \\ &= P(A)P(B^c). \end{aligned}$$

□

Corollary 1.4 *If A and B are independent, then A^c and B^c are independent.*

Pf. If A and B are independent, then A and B^c are independent by Theorem 1.7, but independence is symmetric, so applying Theorem 1.7 again, this time to B^c and A, we conclude B^c and A^c are independent. □

Given a sequence of events $\{A_1, A_2, \ldots, A_n\}$, we say this sequence is independent if the occurrence of any collection of these events does not influence the chance that any other event in the sequence occurs. Formally, it's easiest to state this by extending the relation $P(A \cap B) = P(A)P(B)$ to arbitrary subsets of the sequence.

Definition 1.6 *The sequence of events $\{A_i\}_{i=1}^n$ is independent if for any collection $\{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}$ of events taken from the sequence,*

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_k})$$

Example 1.31 *Suppose we toss a fair coin twice. Let H_1 = 'The first toss is heads', H_2 = 'The second toss is heads', and E = 'Exactly one toss results in heads'. As we already observed, H_1 and H_2 are independent. Moreover, $P(E) = \frac{1}{2}$ since exactly one head is flipped in two of the four equally likely outcomes, and $P(E \mid H_1) = \frac{1}{2}$ since assuming H_1 occurs, E occurs when the second flip is a tail. Thus, H_1 and E are independent, and essentially the same argument shows H_2 and E are independent.*

Since each pair of events in the sequence $\{H_1, H_2, E\}$ is independent, we say this sequence is pairwise independent. However, note that if we assume both H_1 and H_2 occur, the probability E occurs becomes zero, that is, $P(H_1 \cap H_2 \cap E) = 0$, but we can calculate $P(H_1)P(H_2)P(E) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$.

The example above illustrates that even if every pair of events in a sequence is independent, there could still be some event in the sequence whose probability is influenced by the joint occurrence of some larger collection of other events. In other words, for sequences of events, pairwise independence does not imply independence.

Example 1.32 *Suppose that a 64 bit sequence (a sequence of 64 zeros and ones) is sent over an unreliable channel. The probability that each bit gets flipped is 0.005 independent of what happens to the other bits. What is the probability that the entire sequence is received without errors?*

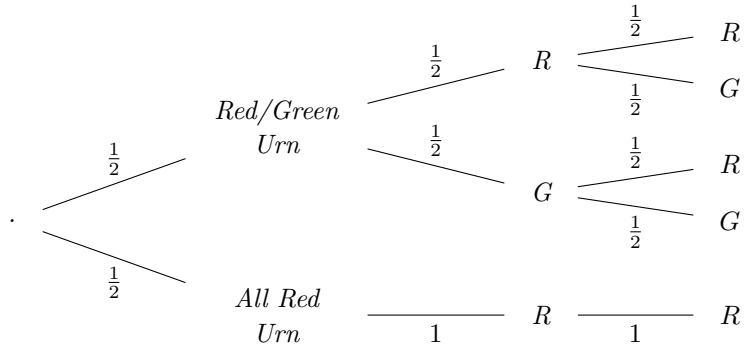
Let F_i = 'the i^{th} bit gets flipped', then $P(F_i) = 0.005$, and the probability that the i^{th} bit is received correctly is $P({F_i}^c) = 1 - 0.005 = 0.995$. If E = 'the entire string is received without any flipped bits' then

$$\begin{aligned} P(E) &= P({F_1}^c \cap {F_2}^c \cap \ldots \cap {F_{64}}^c) \\ &= P({F_1}^c)P({F_2}^c) \ldots P({F_{64}}^c) \\ &= 0.995 \cdot 0.995 \cdot \ldots \cdot 0.995 \\ &\simeq 0.72556 \simeq 72.6\% \end{aligned}$$

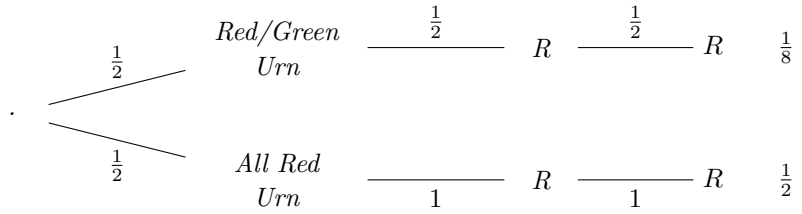
1.7 Bayes' Rule

The tools of conditional probability have so far been used to find the probability of observing a particular outcome under some assumptions about the experiment being performed. In practice though, the problem is often the other way around. We observe a particular outcome, and we'd like to use that information as evidence to evaluate or update our assumptions about the internal workings of the experiment.

Example 1.33 *Two balls are drawn with replacement from an urn which contains either only red balls, or an equal number of red and green balls. The choice of how to load the urn is made by flipping a fair coin. If the experiment is performed, and both balls drawn are red, what is the probability the urn contains only red balls? Consider the tree diagram given below, where every possible outcome is illustrated.*



Now we know that two red balls were drawn, so let's reduce our sample space by disregarding all branches where this event doesn't occur. The updated tree diagram is given below.



Intuitively, since the top branch has a $\frac{1}{8}$ chance of occurring, and the bottom branch has a $\frac{1}{2}$ chance of occurring, the bottom branch is four times more likely to have occurred. Thus, there is a $\frac{4}{5}$ chance that the all red urn was being used, and a $\frac{1}{5}$ chance the urn with an equal number of red and green balls was being used.

Formally, if we let $A =$ 'the urn contains all red balls' then $A^c =$ 'the urn contains and equal number of red and green balls', and we're looking for $P(A \mid R_1 \cap R_2)$.

Using the definition of conditional probability and the law of total probability,

$$\begin{aligned} P(A \mid R_1 \cap R_2) &= \frac{P(A \cap R_1 \cap R_2)}{P(R_1 \cap R_2)} \\ &= \frac{P(A \cap R_1 \cap R_2)}{P(R_1 \cap R_1 \mid A)P(A) + P(R_1 \cap R_2 \mid A^c)P(A^c)} \\ &= \frac{\frac{1}{2} \cdot 1 \cdot 1}{1 \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2}} = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{8}} = \frac{4}{5}. \end{aligned}$$

Initially, the chance the all red urn was selected was $\frac{1}{2}$. This belief was justified by the description of the experiment (whether or not the all red urn was selected was determined by a coin flip). After observing the results of a few draws, we were able to update our beliefs, and the chance the red urn was being used increased to $\frac{4}{5}$. The process of starting from some belief about the state of the world, and updating that beliefs based on new information is, essentially, learning.

The facts we used to justify our conclusion were all derived from the Kolmogorov probability axioms, so the correctness of this approach, which is formalized in Bayes' rule below, is not in question. The fact that we can formalize the process of learning from new information is remarkable, and utility of this result is clear. From automated movie recommendations that change based on your ratings, to self-driving cars that get better at recognizing dangerous situations every time a crash happens, if you want to formalize the process of learning from new data, Bayes' rule is the fundamental tool.

Theorem 1.8 *(Bayes' Rule) For any events A and B in some sample space Ω,*

$$\boxed{P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}}$$

Pf. We can derive Bayes' rule from the definition of conditional probability by applying the intersection rule in the numerator.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B \mid A)}{P(B)}$$

□

It's worth noting that Bayes' rule, the intersection law, and the definition of conditional probability are essentially three different rearrangements of the same statement. These three different ways of thinking about one idea (the concept of conditional probability) lend themselves more naturally to different applications.

Bayes' rule is shown in a compact form above. However in many applications, we will end up splitting the denominator into cases using the law of total probability.

Corollary 1.5 *If $\{A_i\}_{i=1}^n$ is a sequence of events which partitions the sample space Ω, and B is an event,*

$$\boxed{P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B \mid A_1)P(A_1) + \ ... \ + P(B \mid A_n)P(A_n)}}$$

Pf. Apply Bayes' rule as presented in Theorem 1.8 to the events A_i and B, then expand the denominator using the law of total probability. □

This expanded form of Bayes' rule closely matches the reasoning we did with tree diagrams in Example 1.33 above. In the numerator, we have the probability of the branch on which both the event we're assessing the probability of and the evidence we've observed occur, while in the denominator we have the sum of the probabilities of all the branches where the evidence we've observed occurs.

Example 1.34 *Suppose that a rare disease affects* 0.5% *of the population. There is a test for this disease which is* 99% *accurate, in the sense that the probability of a positive test for someone who carries the disease (a true positive) is* 99%*, and probability of a positive test for someone who does not carry the disease (a false positive) is 1%. If an individual is randomly selected from the population and tests positive, what is the probability they have the disease?*

Let $D =$ 'the randomly selected individual tested carries the disease', and $P =$ 'the result of the test is positive'. Using Bayes' rule, we compute

$$\begin{aligned} P(D \mid P) &= \frac{P(P \mid D)P(D)}{P(P \mid D)P(D) + P(P \mid D^c)P(D^c)} \\ &= \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.01 \cdot 0.995} \simeq 0.33. \end{aligned}$$

Thus, if a randomly selected individual tests positive, there's a 33% *chance that they carry the disease. Despite the positive test result, the scenario where this individual does not carry the disease is still the more likely one.*

This example illustrates the problem of over-testing. If data from medical tests is to be useful, they should be applied carefully, often only in cases where there's some reason to believe the patient has an elevated risk of carrying the disease. As the test is more widely applied, the disease becomes more unusual in the group being tested and false positives become more likely. To illustrate this point, consider the most extreme case, where we're testing for a disease which has no cases in the group being tested. Every positive test will be a false positive, regardless of how well-designed the test might be. The utility of any test is a function of not just the intrinsic properties of the test, but also of the nature of the environment in which the test is being applied.

Assessing the usefulness of a procedure by examining only its intrinsic properties and ignoring the context in which it's being applied is known as the *base rate fallacy* and is a very common reasoning error.

Bayes' rule is frequently written using the letters H and E, as below. This is because when employing Bayes' rule, it's often useful to think of E as some evidence which we're using to reassess the probability of a hypothesis H.

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

The quantity $P(H)$ is called a *prior probability* since it represents the probability of the hypothesis before observing the new evidence, while $P(H \mid E)$ is known as a *posterior probability*.

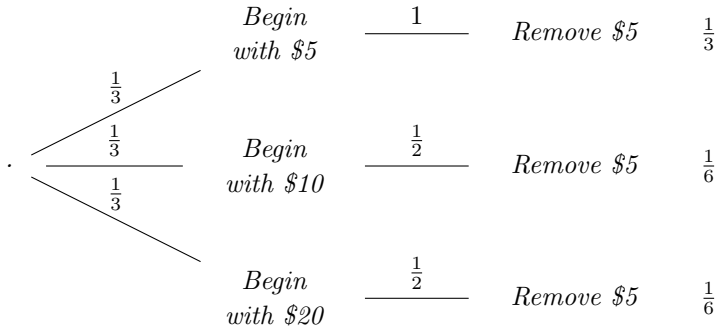
We're now better equipped to understand the Bayesian interpretation of probability given in Section 1.3. Two individuals with the same prior degree of confidence in a hypothesis will agree on what constitutes a rational decision after observing new evidence for that hypothesis, since they will both update their beliefs using Bayes' rule with the same prior probability. However, two individuals who have different prior beliefs may still disagree. This dependence on the knowledge, assumptions, or guesses one has before an experiment occurs means that the Bayesian interpretation of probability applies in cases where neither the classical nor the frequentist interpretation apply, but leaves open the question of how one should choose to assign a prior probability to a given hypothesis when there's no clear choice.

Example 1.35 *Suppose my wallet contains either a \$5 bill, a \$10 bill, or a \$20 bill, with equal likelihood. At some point during the day, I add a \$5 bill. Later in the day, I reach into the wallet and extract a \$5 bill. What is the probability the remaining bill in my wallet is another \$5 bill?*

Let B_i = 'the wallet initially contained an \$i bill' and R_i = 'the bill removed from the wallet was an \$i bill'. Applying Bayes' rule,

$$\begin{aligned} P(B_5 \mid R_5) &= \frac{P(R_5 \mid B_5)P(B_5)}{P(R_5 \mid B_5)P(B_5) + P(R_5 \mid B_{10})P(B_{10}) + P(R_5 \mid B_{20})P(B_{20})} \\ &= \frac{1 \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3}} = \frac{1}{2}. \end{aligned}$$

We could also use a tree diagram to reach the same conclusion, recalling that Bayes' rule includes branches where both the hypothesis and the evidence occur in the numerator, and all branches where the evidence occurs in the denominator. In this case, the evidence consists of removing a \$5 bill from the wallet.



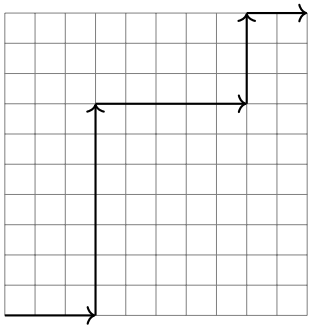
As before, $P(B_5 \mid R_5) = \dfrac{1 \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3}} = \dfrac{1}{2}.$

Chapter 2

Combinatorics

Introduction: Traversing a Grid

Consider a large city laid out in rectangular blocks, and a pedestrian who's walking to his workplace, ten blocks north and ten blocks east of his house. Our pedestrian won't waste his time by ever moving west or south; he has a schedule to keep. How many distinct paths are there from his house to his workplace?



At first glance, he has two options at each intersection he reaches, north or east, but eventually he'll have moved ten blocks in one of those two cardinal directions, and from that point on his decision will be forced. Unfortunately, the number of steps he takes until his options run out is not constant. Some paths will 'hit the wall' sooner than others.

Combinatorics is the field of mathematics concerned with counting problems like this one. These questions are often simple to state, and can have beautifully clever solutions. As we've done with probability theory, we'll develop some general principles which, when carefully applied, provide a basis for building solutions to many counting problems.

2.1 Permutations

Suppose you order onion rings and a soda at a fast food restaurant. There are two options for the size of the onion rings (S/L) while there are three options for the size of the soda ($S/M/L$).

Since there are two options for the size of the onion rings, and three options for the size of the soda, you have a total of six possible choices when you order both items. These are enumerated below, the first letter in each ordered pair denoting the size of the onion rings, and the second letter the size of the soda.

$$(S,S)\ \ (S,M)\ \ (S,L)\ \ (L,S)\ \ (L,M)\ \ (L,L)$$

If there's only enough change in your pocket to pay for a single item (any size), then you have the five options given below.

$$(S,-)\ \ (L,-)\ \ (-,S)\ \ (-,M)\ \ (-,L)$$

Proposition 2.1 *(Fundamental Counting Principle) In a sequence of k choices, if the first choice can be made in n_1 ways, the second choice can be made in n_2 ways, and so on, the number of ways of making all choices is the product $n_1 n_2 \ldots n_k$, while the number of ways of selecting one of the choices and making it is the sum $n_1 + n_2 + \ldots + n_k$.*

Example 2.1 *How many ways are there to select three students to be class representatives in three different classes, one with twenty students and two with thirty?*

There are three choices to be made, choosing a representative for each of the three classes, and since we must make all three choices, the number of ways this can be done is $20 \cdot 30 \cdot 30 = 18\,000$.

Example 2.2 *How many four letter sequences can be made with only the first four letters of the alphabet?*

We can imagine filling in four blank spaces with letters, and write in each blank space the number of choices we have for the letter that gets placed there. If we allow repetition, then the number of sequences is $\underline{4} \cdot \underline{4} \cdot \underline{4} \cdot \underline{4} = 256$.

If each space must be filled with a distinct letter, then initially there are four choices for the first letter, but after making that choice, there are three remaining for the second letter, and so on. In total then, the number of sequences with distinct letters is $\underline{4} \cdot \underline{3} \cdot \underline{2} \cdot \underline{1} = 24$.

This second result is the number of ways of rearranging the four letters $ABCD$, and counting rearrangements is something we'll do frequently, so it's convenient to make some notation.

Definition 2.1 *Given $n \in \mathbb{N}$, the factorial of n, denoted $n!$, is the number of ways of arranging n distinct objects in a line, and is defined by*

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \ ... \ \cdot 2 \cdot 1$$

Remark Though the interpretation of the factorial as counting arrangements no longer applies, we define $0! = 1$ for convenience.

Factorials have some very nice algebraic properties that make them easy to work with, in particular, ratios of factorials can be simplified easily.

$$\frac{8!}{6!} = \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{8 \cdot 7 \cdot \cancel{6} \cdot \cancel{5} \cdot \cancel{4} \cdot \cancel{3} \cdot \cancel{2} \cdot \cancel{1}}{\cancel{6} \cdot \cancel{5} \cdot \cancel{4} \cdot \cancel{3} \cdot \cancel{2} \cdot \cancel{1}} = 56$$

Example 2.3 *How many ways are there to choose three students from a class of twenty and arrange them in a line?*

We can use the same approach as in the last example. We're filling in three blanks, we have twenty options to choose from, and repeats are not permitted. Thus, there are $\underline{20} \cdot \underline{19} \cdot \underline{18} = 6\,840$ ways to do it.

The notation ${}^{20}P_3$ is used to refer to the number of ways of selecting 3 objects from a set of 20 distinct options and ordering them. Thus, we have ${}^{20}P_3 = 6\,840$.

Definition 2.2 *If an ordered list of r distinct objects is selected from a set of n objects, the result is called an r-permutation. The number of r-permutations of n objects, denoted by ${}^nP_r$, is given by*

$$\boxed{{}^nP_r = n \cdot (n-1) \cdot \ ... \ \cdot (n-r+1) = \frac{n!}{(n-r)!}}$$

Example 2.4 *How many ways there there to create a sequence of five letters with no repetitions using any of the 26 letters of the alphabet?*

Taking 5 letters from a set of 26 and arranging them, the number of options is

$${}^{26}P_5 = \frac{26!}{(26-5)!} = \frac{26!}{21!} = 26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 = 7\,893\,600$$

Permutations with Identical Objects

What if some objects are indistinguishable? If this is the case, there are fewer possibilities. For example, how many arrangements of the string $AAAB$ are there? We just have to decide which of the four possible positions we'll place the B into, and then As go everywhere else, so there are only four distinct arrangements.

Now suppose that we number the three As. How many rearrangements of the string $A_1A_2A_3B$ are there? Since the four letters are now distinguishable, there are

$4! = 24$. Let's write these out, grouping together in columns all strings that will be identical after the labels on the As are removed.

$$\begin{array}{cccc} A_1A_2A_3B & A_1A_2BA_3 & A_1BA_2A_3 & BA_1A_2A_3 \\ A_1A_3A_2B & A_1A_3BA_2 & A_1BA_3A_2 & BA_1A_3A_2 \\ A_2A_1A_3B & A_2A_1BA_3 & A_2BA_1A_3 & BA_2A_1A_3 \\ A_2A_3A_1B & A_2A_3BA_1 & A_2BA_3A_1 & BA_2A_3A_1 \\ A_3A_1A_2B & A_3A_1BA_2 & A_3BA_1A_2 & BA_3A_1A_2 \\ A_3A_2A_1B & A_3A_2B_1A & A_3BA_2A_1 & BA_3A_2A_1 \end{array}$$

Notice that rearranging the labels will result in a string which is identical once the labels are removed. This means if we list all possible arrangements of the *labeled strings*, each possible *unlabeled string* will appear $k!$ times, where k is the number of labels. In this case, each distinct unlabeled string appears $3! = 6$ times.

The key point is that we can count the number of unlabeled strings by dividing the number of labeled strings by the number of ways of rearranging the labels, in this case, $4!/3! = 4$. The same principle applies if there are many different kinds of indistinguishable objects.

Proposition 2.2 *(Mississippi Formula) The number of ways of rearranging n objects, with indistinguishable groups of sizes $k_1, k_2, \ldots, k_m$, is given by*

$$\boxed{\frac{n!}{k_1! \cdot k_2! \cdot \ \ldots \ \cdot k_m!}}$$

Example 2.5 *How many arrangements of the letters in MISSISSIPPI are there?*

There are eleven letters in total, with identical groups of size 4 (the I's), 4 (the S's), and 2 (the P's). Therefore, the number of arrangements is

$$\frac{11!}{4! \cdot 4! \cdot 2!} = \frac{11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 2 \cdot 1} = 11 \cdot 10 \cdot 9 \cdot 7 \cdot 5 = 34\,650.$$

Example 2.6 *An urn contains four green and five red marbles, which are removed one at a time. How many different ways are there to empty the urn?*

In this case, we are rearranging nine objects, each arrangement corresponds to a possible order in which they might appear when removed from the urn. There are identical groups of size 4 (the green marbles), and 5 (the red marbles). Therefore, the number of arrangements is

$$\frac{9!}{4! \cdot 5!} = \frac{9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 9 \cdot 2 \cdot 7 = 126.$$

2.2 Combinations

Suppose now that we are interested in the number of ways that a subset of r objects can be selected from a larger set of n distinct objects. The order in which

the selections are made doesn't matter, we're only concerned with which objects are selected.

Consider how many ways there are to select three children from a family with five. To do this, we can imagine lining up the children any fixed order, let's say by age. We can now specify which children are chosen using a string such as $YYNYN$, which would indicate we chose the first two children in the line as well as the fourth. In fact, every distinct five letter string made up of Ys and Ns which has exactly three Ys will represent a different choice of three children, and every choice of three children determines such a string.

Now we can use the Mississippi formula to count the number of different choices by counting strings. Each string is a rearrangement of five objects with identical groups of sizes three and two, hence there are $\frac{5!}{3!2!} = 5 \cdot 2 = 10$ possibilities.

The general principle is that once the elements of a set have been lined up in some arbitrary order, each subset is given by an *ordered* list of Y's and N's, where Y denotes membership in the subset, and N denotes non-membership. We can count the number of ordered lists of Y's and N's by using the MISSISSIPPI formula.

Definition 2.3 *If a subset of r objects is selected from a set of n objects, the result is called an r-combination. The number of r-combinations, denoted ${}^nC_r$, or more commonly $\binom{n}{r}$, is given by*

$$\boxed{{}^nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}}$$

Remark Note the distinction between ${}^nP_r$ and ${}^nC_r$. The former counts the number of ways of selecting r objects *and arranging them*, while the latter is the number of selections possible when *any rearrangement is regarded as the same outcome*.

Example 2.7 *In Lotto 6/49, if you purchase a ticket you choose 6 distinct numbers between 1 and 49. The ticket wins if its numbers match 6 distinct numbers between 1 and 49 which are randomly drawn, regardless of the order they are drawn in. How many different tickets would you need to buy to guarantee a win?*

Since the order of the numbers on the ticket is irrelevant, the number of different tickets available is $\binom{49}{6} = \frac{49!}{6!43!} = 13\,983\,816$, so this is the minimum number you would need to purchase to be sure one will win.

Example 2.8 *How many ways are there to choose a president, a treasurer, and an organizing committee of three from a group of 25 people? Assume no person can hold more than one position.*

There are 25 choices for president, then 24 for treasurer, then $\binom{23}{3}$ for the organizing committee. Note that when choosing the organizing committee, the order in which members are chosen is irrelevant. The committee consisting of Alice, Bob, and Chelsea and the committee consisting of Chelsea, Alice, and Bob are the same.

Since we must make all three choices, by the fundamental counting principle the number of possibilities is

$$25 \cdot 24 \cdot \binom{23}{3} = 1\,062\,600.$$

Notice that we could just as well have proceeded in the opposite order in our analysis, that is, first chosen the organizing committee, then the treasurer, then the president, and still obtained

$$\binom{25}{3} \cdot 22 \cdot 21 = 1\,062\,600.$$

Example 2.9 *How many ways are there to divide a group of eight people into a team of three and a team of five?*

To create a team of three and a team of five, we choose three individuals for the team of three, then the five remaining make up the team of five.

$$\binom{8}{3}\binom{5}{5} = 56$$

Example 2.10 *How many ways are there to divide a group of eight people into two teams of four?*

To create two teams of four, we choose four for the first team, and those that remain make up another team of four. We have to be very careful though. In the example above the two teams were distinguishable, but here they are not.

If we select A, B, C, and D for the first team and E, F, G, and H for the second, this results in the outcome as selecting E, F, G, and H for the first team and A, B, C, and D for the second. The teams are indistinguishable, so every outcome has been double counted! Thus, we need to divide out by the number of rearrangements of the teams.

$$\frac{\binom{8}{4}\binom{4}{4}}{2!} = 35$$

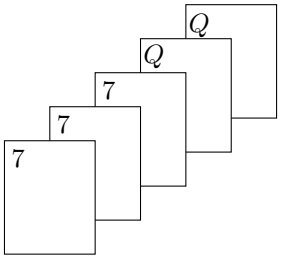
Notice that if the two teams were distinguishable, for example, if we divide eight people into a marketing team and an administrative team with four people on each team, there would be $\binom{8}{4}\binom{4}{4} = 70$ *ways.*

Poker Hand Probabilities

In many variations of the gambling game poker, five cards are drawn from a shuffled deck to form a hand. The order of these cards is irrelevant, each player may freely rearrange his or her hand. We can use the counting techniques we've developed to compute the probabilities of drawing various hands.

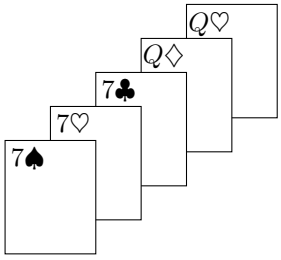
Example 2.11 *What is the probability of drawing a full house? In this hand, three of the five cards share a rank, and the remaining two also share a rank, for example, the three cards could be* $7\spadesuit\, 7\heartsuit\, 7\clubsuit\, Q\diamondsuit\, Q\heartsuit$.

There are 13 possible ranks for the group of three cards and we must choose one. Once that's done, there is a pool of 12 remaining ranks for the other two, from which we must choose one. Thus, to fill in the ranks, we have $\binom{13}{1}\binom{12}{1}$ *possibilities. If we choose, for example,* 7 *for the group of three and* Q *for the remaining two, we've determined the ranks as shown below.*



Note that if we instead choose Q *for the group of three and* 7 *for the remaining two, we would produce a different hand, which has three queens and two sevens. This is good. We need distinct sequences of choices to produce distinct hands. If not, we would double count certain hands.*

To determine the suits, notice that we have four choices for each card, but within each group of identical ranks, suits cannot be repeated (there is only one 7 of hearts in a deck), and order does not matter since we can freely reorder the hand. Thus, there are $\binom{4}{3}$ *options for the suits in the group of three cards of equal rank, and* $\binom{4}{2}$ *options for the suits in the group of two. Since both choices need to be made, we have a total of* $\binom{4}{3}\binom{4}{2}$ *options when determining the suits. After filling in the suits, we have completely determined the hand.*

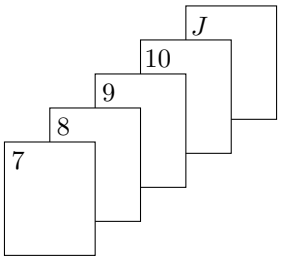


Thus, there are $\binom{13}{1}\binom{12}{1}\binom{4}{3}\binom{4}{2}$ *different full houses. Since there are* $\binom{52}{5}$ *possible five card subsets of a standard deck, and each is equally likely to be drawn if we suppose the deck is well shuffled, the probability of a full house is*

$$\frac{\binom{13}{1}\binom{12}{1}\binom{4}{3}\binom{4}{2}}{\binom{52}{5}} = \frac{3\,744}{2\,598\,960} \simeq 0.14\%.$$

Example 2.12 *What is the probability of drawing a straight? In this hand, all five cards have consecutive ranks, for example, the three cards could be* 7♠ 8♢ 9♡ 10♡ J♢*. Note that A plays the role of both the highest and lowest possible rank.*

Again, let's deal with the ranks first, and then the suits. In a straight, the card of lowest rank determines the ranks of all other cards. If the lowest ranked card is a 7*, then each rank between* 7 *and J will appear once in the hand, as shown below.*



Since there are 10 *choices for the lowest ranked card, A through* 10*, and all the rest of the ranks are determined by that single card, there are* 10 *possibilities. Now that the ranks are dealt with, what about the suits? Each card has a different rank, so suits can be repeated. There are* 4 *choices for the suit of each card, and five cards, hence* 4^5 *ways to assign suits by the fundamental counting rule.*

In total then, there are $10 \cdot 4^5$ *possible straights, and thus the probability of drawing a straight is*

$$\frac{10 \cdot 4^5}{\binom{52}{5}} = \frac{10\,240}{2\,598\,960} = 0.39\%.$$

Note that if all five cards also have the same suit, the hand is called a straight flush, so here we're including straight flushes (and royal flushes). If you'd like to correct for this, find the number of straight flushes (and royal flushes) and subtract it out.

2.3 The Binomial Theorem

Consider the expression $(x + y)^n$, where $n \in \mathbb{N}$. Let's take the first few natural numbers for n and expand this expression.

$$(x+y)^1 = x + y$$

$$\begin{aligned}(x+y)^2 &= (x+y)(x+y)\\ &= xx + xy + yx + yy\\ &= x^2 + 2xy + y^2\end{aligned}$$

$$\begin{aligned}(x+y)^3 &= (x+y)(x+y)(x+y)\\ &= xxx + xxy + xyx + xyy + yxx + yxy + yyx + yyy\\ &= x^3 + 3x^2y + 3xy^2 + y^3\end{aligned}$$

Once the exponent gets much larger, it becomes a real headache to go through the ordeal of expanding everything out. Wouldn't it be nice if there was a clever way to get straight to the result without having to go through all the work of expanding and collecting like terms?

To expand $(x+y)(x+y)(x+y)$, we need to select a term from each of the three copies of the binomial expression $x+y$, either the left term, or the right term. We do this in each possible way one time to obtain the expansion. For instance, if we take the left term in the first two factors, and the right in the last,

$$(\underline{x}+y)(\underline{x}+y)(x+\underline{y}) \;\rightarrow\; xxy.$$

In this way, we can associate strings of three L's and R's to terms in the expansion. Every such string will yield a distinct term in the expansion, simply by identifying L with x and R with y.

$$\begin{aligned} LLR &\;\rightarrow\; (\underline{x}+y)(\underline{x}+y)(x+\underline{y}) \;\rightarrow\; xxy \\ RRR &\;\rightarrow\; (x+\underline{y})(x+\underline{y})(x+\underline{y}) \;\rightarrow\; yyy \\ RLR &\;\rightarrow\; (x+\underline{y})(\underline{x}+y)(x+\underline{y}) \;\rightarrow\; yxy \end{aligned}$$

Consider the $3x^2y$ term in the expression $x^3+3x^2y+3xy^2+y^3$. It was obtained by adding up all terms in the expansion which had two x's and one y. Each of these corresponds to a string with two L's and one R. There are $\frac{3!}{2!1!}=3$ such strings by the Mississippi formula, and hence the coefficient of x^2y after expanding $(x+y)^3$ and collecting like terms is 3. The same idea applies in general.

The coefficient of $x^k y^{n-k}$ in the expansion of $(x+y)^n$ is the same as the number of strings of L's and R's of length n with exactly k L's, which is $\frac{n!}{k!(n-k)!}=\binom{n}{k}$ by the Mississippi formula.

This idea is often stated in the form of the theorem given below, and values of the form $\binom{n}{k}$ for some n and k are often referred to as binomial coefficients because of this connection.

Theorem 2.1 *(Binomial Theorem)*

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^{n-i}y^i = \binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + ... + \binom{n}{n}y^n$$

Remark The first and last coefficients in the expansion are both one. This provides some motivation for defining $0!=1$; it makes $\binom{n}{0}$ and $\binom{n}{n}$ both defined and equal to one, so the binomial theorem is simpler to state.

Example 2.13 *What is the last digit of the number* 109^{2717}*?*

This number is much too large to actually write down digit by digit, but with clever use of the binomial theorem, we can determine the last digit without knowing any of the others. Notice that $109^{2717}=(110-1)^{2717}$*, and by the binomial theorem,*

$$(110-1)^{2717} = \binom{2717}{0}110^{2717} + \binom{2717}{1}110^{2716}\cdot(-1)^1 + ... + \binom{2717}{2717}(-1)^{2717}$$

Each term is a product of integers, and 110 *is a multiple of ten, hence so is* 110^k *for any integer* k*. Therefore, every term in the expansion is a multiple of ten except the last. This final term is* $(-1)^{2717} = -1$*. Therefore,* 109^{2717} *is one less than a multiple of ten, hence its last digit is a 9.*

This argument can be generalized considerably. Take a power of any number, write the base as a sum or difference of a multiple of ten and a single digit number, then use the binomial theorem. If you can compute the last digit of the last term of the expansion, you'll know the last digit of the number in question.

Pascal's Triangle

$$
\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
 & \vdots & & & & & & & & \vdots &
\end{array}
$$

This unending arrangement of numbers is commonly known as Pascal's triangle, in honour of Blaise Pascal, French mathematician and founding father of probability theory. Each row begins and ends with a one, while all interior entries are obtained by summing the two entries on the left and right in the row above. For instance, if we were to write the next row, 70 would appear in the center.

Observe that the second, third, and fourth rows are exactly the coefficients that appeared in the expansions of $(x+y)^1$, $(x+y)^2$, and $(x+y)^3$. Could this be true in general? Why?

Amazingly enough, it's true! The numbers that appear in the $(n+1)^{th}$ row of Pascal's triangle are exactly the coefficients in the expansion of $(x+y)^n$. This can be a neat shortcut when we apply the binomial theorem. To expand $(a+x)^6$, for example, write the first seven rows of Pascal's triangle. From the seventh row,

$$(a+x)^6 = a^6 + 6a^5x + 15a^4x^2 + 20a^3x^3 + 15a^2x^4 + 6ax^5 + x^6.$$

To understand how this can be, let's play with Pascal's triangle a little. Consider a sequence of steps, starting from the top of the triangle, each step moving either down & left, or down & right. Let's call such a sequence of steps a path. Take, for example, the highlighted path below which ends in the sixth row.

$$
\begin{array}{ccccccccccc}
 & & & & & \mathbf{1} & & & & & \\
 & & & & \mathbf{1} & & 1 & & & & \\
 & & & 1 & & \mathbf{2} & & 1 & & & \\
 & & 1 & & 3 & & \mathbf{3} & & 1 & & \\
 & 1 & & 4 & & \mathbf{6} & & 4 & & 1 & \\
1 & & 5 & & \mathbf{10} & & 10 & & 5 & & 1 \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \\
1 & & 7 & & 21 & & 35 & & 35 & & 21 & & 7 & & 1 \\
 & & & & & \vdots & & & & &
\end{array}
$$

As it meanders down the triangle, this path moved left, then right, then right, then left, then left, so it can be denoted using the string LRRLL. In fact, every string of five L's and R's with three L's describes a path which terminates in the same place. For instance, RLLLR gives the path highlighted below.

$$
\begin{array}{ccccccccccc}
 & & & & & \mathbf{1} & & & & & \\
 & & & & 1 & & \mathbf{1} & & & & \\
 & & & 1 & & \mathbf{2} & & 1 & & & \\
 & & 1 & & \mathbf{3} & & 3 & & 1 & & \\
 & 1 & & \mathbf{4} & & 6 & & 4 & & 1 & \\
1 & & 5 & & \mathbf{10} & & 10 & & 5 & & 1 \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \\
1 & & 7 & & 21 & & 35 & & 35 & & 21 & & 7 & & 1 \\
 & & & & & \vdots & & & & &
\end{array}
$$

We already discovered that the number of strings with five L's and R's containing exactly three L's is $\binom{5}{3} = 10$, but how does this same number miraculously appear when instead of counting paths, we simply add the 4 and 6 in the row above?

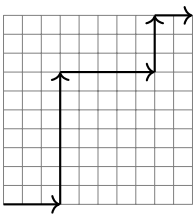
The key point is that *every path which ends at a particular location is obtained either by adding an R to a path which ended directly above and to the left, or by adding an L to a path that ended directly above and to the right*. Thus, the number of paths that end at any given location is always the sum of the number of paths that end directly above and to the left and the number of paths that end directly above and to the right.

Proposition 2.3 *The number appearing $k+1$ places across in row $n+1$ of Pascal's triangle is $\binom{n}{k}$, and furthermore, this number counts each of the following.*

- *Subsets containing k objects taken from a collection of n distinct objects.*
- *The coefficient of $x^{n-k}y^k$ in the expansion of $(x+y)^n$.*
- *Strings of L's and R's of length n with exactly k L's.*
- *Paths through Pascal's triangle ending $k+1$ places across in row $n+1$.*

The problem presented at the beginning of this chapter can be solved using the same idea of representing paths with strings and counting the number of strings.

Example 2.14 *Consider a large city laid out in rectangular blocks, and a pedestrian who's walking to his workplace, ten blocks north and ten blocks east of his house. Our pedestrian won't waste his time by ever moving west or south; he has a schedule to keep. How many distinct paths are there from his house to his workplace?*



Each path can be represented with a string of E's and N's, representing a sequence of movements one unit east (E), or one unit north (N). The path shown above, for example, corresponds to the string $EEENNNNNNNEEEEENNNEE$. Every path yields a string of 20 *E's and N's, exactly 10 of which are E's. Therefore, the total number of possible paths is*

$$\binom{20}{10} = 184\,756.$$

2.4 Partitions

Suppose we have 7 identical marbles and 3 distinguishable boxes. How many ways are there to partition the marbles among the boxes?

There's a really beautiful and simple way to solve this problem. We represent the seven marbles with seven identical stars. To partition the marbles among the boxes, we simply insert two bars. The diagram below, for example, represents a partition with one marble in the first box, two marbles in the second, and four marbles in the third.

$$*\,|\,**\,|\,****$$

To represent a partition where all marbles are in the first box, we can draw the following diagram.

$$*******\,||$$

If all marbles are placed in the third box, we have the diagram below. Note that this is a different partition than the last, since the three boxes are distinguishable.

$$||\,*******$$

This idea, often called the stars & bars method, was popularized in a book by William Feller [2]. To count the number of partitions, we count the diagrams.

Each diagram consists of 9 symbols, 2 of which are bars. Once the positions of the bars have been determined, all other positions must be filled in with stars. The order of the choices is not relevant, only which positions are chosen for the bars. Thus, there are $\binom{9}{2} = 36$ ways to partition the 7 marbles among the 3 boxes.

Proposition 2.4 *(Stars & Bars) The number of ways of partitioning n identical objects into k distinguishable groups is*

$$\binom{n+k-1}{k-1}$$

Example 2.15 *How many ordered triples (x, y, z) of nonnegative whole numbers are solutions to the equation $x + y + z = 15$?*

In this case, we can think of 15 as the sum of 15 ones, and partition these ones into three groups, one for each variable. The stars & bars diagram below, for example, represents the partition $7 + 3 + 5$.

$$* * * * * * * | * * * | * * * * *$$

Thus, there are $\binom{17}{2} = 136$ ordered triples which sum to 15.

Example 2.16 *How many ordered triples (x, y, z) of positive whole numbers are solutions to the equation $x + y + z = 15$?*

Notice the difference. Zero is not a positive number, so we are not allowing empty groups. We would like to exclude partitions such as $7 + 9 + 0$.

$$* * * * * * * | * * * * * * * * * |$$

There is a very simple yet very clever idea we can use. Notice that if $x+y+z = 12$, then $(x + 1) + (y + 1) + (z + 1) = 15$. Every ordered triple of nonnegative numbers which sums to 12 will yield a unique ordered triple of strictly positive numbers that sums to 15 after we add one to each variable.

In other words, we can imagine that before the partitioning, each variable already starts at one, so we partition the 12 remaining ones to end up with a sum of 15. Thus, there are $\binom{14}{2} = 91$ possibilities.

Example 2.17 *How many positive whole numbers under 10 000 have a digit sum of twelve? (For example, 7014 has a digit sum of twelve since $7 + 0 + 1 + 4 = 12$)*

All such numbers have four or fewer digits, and if we allow digits to be zero, then we're simply partitioning twelve ones into four groups, one for each digit. As above, we can represent such a partition using stars & bars. The number 3063, for example, corresponds to the diagram below.

$$* * * | | * * * * * * | * * *$$

However, there is a problem. Some partitions will place more than 9 stars between two bars, and the largest possible digit is 9. We'll have to subtract these to come up with the correct count.

Suppose that the first group (the leftmost) contains more than nine stars. We can generate all such diagrams by assuming ten stars have already been placed in the first group, then partitioning the rest freely. Since there are two stars left to partition using the three bars, there are a total of $\binom{5}{2}$ *possibilities.*

The same argument can be made for diagrams with more than nine stars in each of the four possible groups, hence there are total of $4 \cdot \binom{5}{2}$ *partitions we need to subtract. Therefore, the number of positive whole numbers under 10 000 whose digits sum to twelve is* $\binom{15}{3} - 4 \cdot \binom{5}{2} = 415$.

Chapter 3

Discrete Random Variables

Introduction: Payout on Tails

Consider a gambling game in which a coin is repeatedly flipped until the first tail appears. The payout is determined by the number of heads that occur before this first tail. If n heads occurred then the player receives $\$\, n$.

$$\begin{aligned} T &\to \$\,0 \\ HT &\to \$\,1 \\ HHT &\to \$\,2 \\ HHHT &\to \$\,3 \\ &\vdots \end{aligned}$$

How much should a rational player be willing to pay in advance to play this game? Certainly if you could play for free you would, and it's not hard to convince yourself that paying $50\,¢$ to play is a rational decision. Why? The first flip is heads with probability $\frac{1}{2}$, and in this case the payout will be at least $\$\,1$, possibly much more. Hence the two events 'leaving with $50\,¢$ less than you came with' and 'leaving with at least $50\,¢$ more than you came with' are equally likely.

We would like to be more precise, and determine exactly where the break even point lies. Is it still rational to play the game for $\$\,0.75$? For $\$\,1$? For $\$\,2$? At what point does the game shift from favouring the player to favouring the house?

The player's payout in this game is an example of a *random variable*, which is a real number whose value is determined by the outcome of an experiment that involves some randomness or element of chance. Random variables are the central object of study in probability & statistics, and they'll stick around for the rest of the course.

3.1 Discrete Probability Distributions

Definition 3.1 *A random variable is a function from a sample space Ω to $\mathbb{R}$.*

If all values in the range of this function can be listed one-by-one (whether that list is finite or infinite), the random variable is called discrete. In this section, we'll deal only with discrete random variables.

Recall that a sample space represents the set of all outcomes of some experiment, so a random variable associates a real number with each possible outcome. In other words, a random variable is simply a variable whose value depends on the outcome of an experiment.

Random variables are typically denoted using capital letters at the end of the alphabet, such as X, Y, or Z, and the probability the random variable X takes the value a is denoted $P(X = a)$. This probability can be determined by computing the probability of an event in the sample space on which X was defined.

Example 3.1 *Let X denote the outcome of single die roll. Then X is a random variable defined on the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$.*

Not surprisingly, $P(X = 1) = \frac{1}{6}$, $P(X = 2) = \frac{1}{6}$, and so on. We could also ask for $P(X > 4)$, and again you should not be surprised that $P(X > 4) = \frac{2}{6}$.

Notice that any equation or inequality involving a random variable determines some subset of outcomes (an event) in Ω, and so the probability of the equation or inequality holding is the probability that this event occurs.

In fact, for any event $E \subseteq \Omega$, we can define an indicator variable I_E, which takes the value 1 if E occurs and 0 if E^c occurs.

Example 3.2 *Consider rolling a pair of dice and define $S =$ 'a sum of seven was rolled', find $P(I_S = 1)$ and $P(I_S = 0)$.*

By definition of I_S, we have $P(I_s = 1) = P(S)$ and $P(I_s = 0) = P(S^c)$. From the table constructed in Section 1.1, $P(I_s = 1) = \frac{6}{36}$ and $P(I_s = 0) = \frac{30}{36}$.

Example 3.3 *Let X denote the number of hearts in three draws, with replacement, from a shuffled deck. What is $P(X > 1)$?*

We are asking for the probability that more than one heart is drawn in three draws, with replacement. Using the law of total probability or a tree diagram,

$$P(X > 1) = \frac{1}{64} + \frac{3}{64} + \frac{3}{64} + \frac{3}{64} = \frac{10}{64} \simeq 0.156.$$

Remark Note that random variables must take real number values. If we let X denote the number of heads in a single flip of a coin, then X is a random variable,

but if we let X denote the outcome (Heads or Tails) of a single flip of a coin, then X is not a random variable.

We can summarize a random variable by listing every possible value it can take, and writing the probability it takes each of those values. To do so is to give the probability distribution of the random variable X.

Definition 3.2 *The probability distribution of a random variable X associates to every possible value of X the probability that it occurs.*

Often we'll drop the word probability from the definition and refer simply to 'the distribution of X'. The simplest way to present the distribution of a discrete random variable is through a table.

Example 3.4 *Suppose that a fair coin is flipped twice, and let X be the number of heads observed. Write the distribution of X using a table.*

Let's take $\Omega = \{HH, HT, TH, TT\}$ for our sample space, and since the coin is fair, each of the four outcomes should be equally likely. In two of the four outcomes, one head is observed, in one outcome no heads are observed, and in one outcome two heads are observed.

x	$P(X = x)$
0	$\frac{1}{4}$
1	$\frac{2}{4}$
2	$\frac{1}{4}$

Remark Notice that capital letters are used to denote random variables while lower case letters are used to denoted possible values that random variables may take.

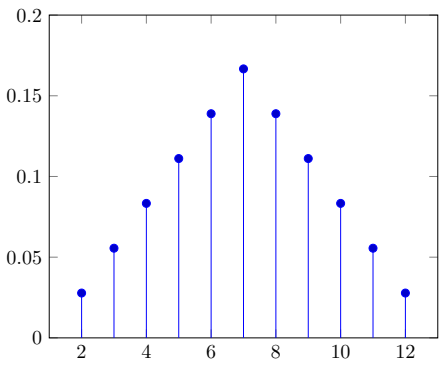
Probability Mass Functions

If we define a function $f_X : \mathbb{R} \to \mathbb{R}$ by assigning to each possible value of X the probability it occurs, the result is called the probability mass function of X. Note that if x is any value which X never takes, then we define $f_X(x) = 0$.

Definition 3.3 *The probability mass function (abbreviated pmf) f_X for the random variable X is defined by $f_X(x) = P(X = x)$.*

Example 3.5 *Let Y denote the sum of the to numbers showing when a pair of dice is rolled. Graph the probability mass function f_Y.*

From the table in Section 1.1, we know Y can take any integer value between 2 and 12 inclusive. We can also see $P(Y = 2) = \frac{1}{36}$, $P(Y = 3) = \frac{2}{36}$, $P(Y = 4) = \frac{3}{36}$, and so on. Plotting these values, we obtain the probability mass function f_Y shown below.



The vertical lines under each point are not technically part of the graph of f_Y, but are often drawn to make the graph easier to construct and interpret. One can also think about the pmf as a bar graph displaying the number of outcomes for each possible value of X when we take a very large number of samples.

Proposition 3.1 *Suppose that X is a discrete random variable which takes values in the set $A = \{x_1, x_2, x_3, ...\}$. Then the probability mass function f_X satisfies the following properties.*

(i) $f_X(x) \geq 0$ *for all* $x \in \mathbb{R}$.

(ii) $f_X(x) = 0$ *for all* $x \notin A$.

(iii) $\sum_{x \in A} f_X(x) = 1$.

In fact, these properties completely characterize the class of probability mass functions for discrete random variables. In other words, if f is a function with these three properties, then it is the pmf of a discrete random variable.

Example 3.6 *Find the value of k so that the function h below is a pmf.*

$$h(x) = \begin{cases} \left(\frac{1}{3k}\right)^x & \text{if } x = 1, 2, 3, ... \\ 0 & \text{otherwise} \end{cases}$$

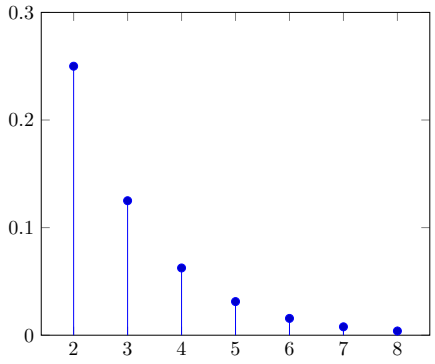
As long as k is positive, h will satisfy properties (i) and (ii) above. The last property allows us to determine k uniquely.

$$\sum_{x \in A} h(x) = \sum_{x=1}^{\infty} \left(\frac{1}{3k}\right)^x = \frac{1}{3k} + \left(\frac{1}{3k}\right)^2 + \left(\frac{1}{3k}\right)^3 + ... = \frac{\frac{1}{3k}}{1 - \frac{1}{3k}} = \frac{1}{3k - 1}$$

The geometric series sum formulas was used to evaluate the infinite sum. Setting the last expression equal to 1 gives $3k - 1 = 1$, and hence $k = \frac{2}{3}$. Using this value of k gives the function h below, which we can graph as a pmf.

$$h(x) = \begin{cases} \left(\frac{1}{2}\right)^x & \text{if } x = 1, 2, 3, ... \\ 0 & \text{otherwise} \end{cases}$$

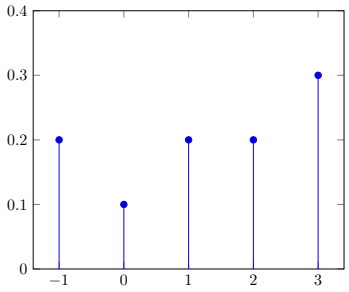
Of course, only the first few values can be shown, since the graph continues off to the right indefinitely.



Functions of a Discrete Random Variable

If X is a discrete random variable, and $g : \mathbb{R} \to \mathbb{R}$ is any function, then $Y = g(X)$ is also a discrete random variable. Its value is completely determined by X.

Example 3.7 *Let X be the random variable whose pmf is given below, and let $Y = 1 + X^2$. Find the pmf of Y.*

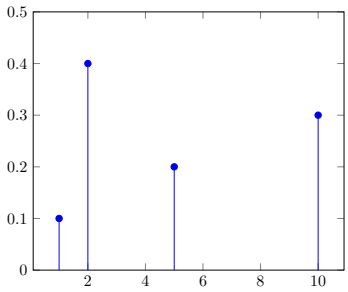


We need to calculate Y for each possible value of X, and sum up the probabilities when appropriate. In this case, note that there are two values of X which result correspond to $Y = 2$, so $f_Y(2) = f_X(-1) + f_X(1)$.

x	y	$f_X(x)$
-1	2	0.2
0	1	0.1
1	2	0.2
2	5	0.2
3	10	0.3

Now we know every possible value of Y and the corresponding probability of each, so we can graph the pmf of Y.

y	$f_Y(y)$
1	0.1
2	0.4
5	0.2
10	0.3



Cumulative Distribution Functions

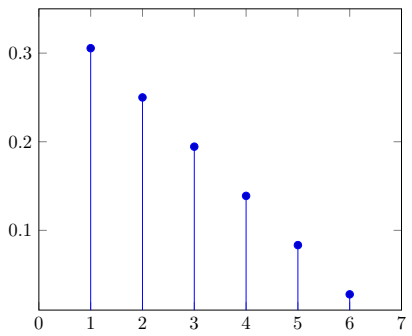
Definition 3.4 *The cumulative distribution function (abbreviated cdf) F_X for the random variable X is defined by $F_X(x) = P(X \leq x)$.*

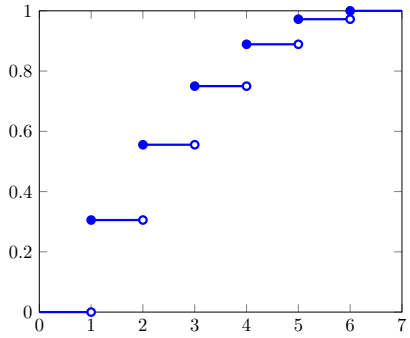
Notice that we use capital letters for the cdfs and lower case letters for pmfs. This is conventional and we'll always do it. It's also worth mentioning that the word 'cumulative' is frequently dropped, and the cdf is called the distribution function.

Example 3.8 *Suppose we roll a pair of dice, and let X be the minimum of the two values that appear on the dice. Find the pmf f_X and the cdf F_X.*

The random variable X takes integer values between 1 and 6 inclusive, and referring back to the table in Section 1.1, we can find the probability X takes the value x by counting the number of cells where the smaller of the two values rolled is equal to x. We have $P(X = 1) = \frac{11}{36}$, $P(X = 2) = \frac{9}{36}$, $P(X = 3) = \frac{7}{36}$, and so on.

The graph of the pmf f_X is shown on the left, and the cdf F_X is shown on the right.





Notice that $F_X(2.5) = P(X \leq 2.5) = P(X = 1) + P(X = 2) = \frac{11}{36} + \frac{9}{36} \simeq 0.56$. The cumulative distribution function adds up the probabilities that X takes any value smaller than or equal to its input.

Consider two real numbers a and b with $a \leq b$. Given any random variable X, $F_X(a) \leq F_X(b)$, because if X is smaller than a then it must also be smaller than b, in other words, the event $X \leq a$ is a subset of the event $X \leq b$, and hence it cannot be more likely. This means that a cdf is always non-decreasing. This and other key properties that all cdfs share are listed in the proposition below.

Proposition 3.2 *For any random variable X, the distribution function F_X satisfies the following properties.*

(i) $F_X(x) \geq 0$ *for all* $x \in \mathbb{R}$.

(ii) F_X *is non-decreasing.*

(iii) $\lim_{x \to -\infty} F_X(x) = 0$ *and* $\lim_{x \to \infty} F_X(x) = 1$.

(iv) $F_X(x)$ *is right-continuous, meaning* $\lim_{x \to a^+} F_X(x) = F_X(a)$.

In fact, these properties completely characterize distribution functions, that is, any function satisfying all four properties is the cdf of some random variable. You should be able to justify the first three properties from the definition of a cdf, but it's not immediately clear clear why we might expect the fourth property to hold.

Theorem 3.1 *If X is any random variable, then its distribution function F_X is right-continuous.*

Pf. Take any $a \in \mathbb{R}$, and consider a decreasing sequence of real numbers $\{x_i\}_{i=1}^{\infty}$ that converges to a. If you'd like a concrete example to think about, consider the sequence $a+1$, $a+\frac{1}{2}$, $a+\frac{1}{4}$, $a+\frac{1}{8}$, and so on.

Define events $\{A_i\}_{i=1}^{\infty}$ from this sequence by letting A_i be the event 'X is less than x_i'. Since the sequence $\{x_i\}_{i=1}^{\infty}$ is decreasing, $\{A_i\}_{i=1}^{\infty}$ is a decreasing sequence of events, so by Theorem 1.5,

$$P(\cap_{i=1}^{\infty} A_i) = \lim_{i \to \infty} P(A_i).$$

But the event $\cap_{i=1}^{\infty} A_i$ occurs when all A_i occur, which means that X is smaller than any element of the sequence $\{x_i\}_{i=1}^{\infty}$, and this means $X \leq a$. Therefore we have

$$P(X \leq a) = \lim_{i \to \infty} P(A_i).$$

Now $P(X \leq a) = F_X(a)$ by definition of the cdf of X, and since A_i occurs when $X \leq x_i$, we have $P(A_i) = P(X \leq x_i) = F_X(x_i)$ as well. Thus,

$$F_X(a) = \lim_{i \to \infty} F_X(x_i).$$

Since $\{x_i\}_{i=1}^{\infty}$ was an arbitrary decreasing sequence that converges to a, the limits $\lim_{i \to \infty} F_X(x_i)$ and $\lim_{x \to a^+} F_X(x)$ are equal (in both cases we are approaching a from the right) so we have the desired equality.

$$F_X(a) = \lim_{x \to a^+} F_X(a).$$

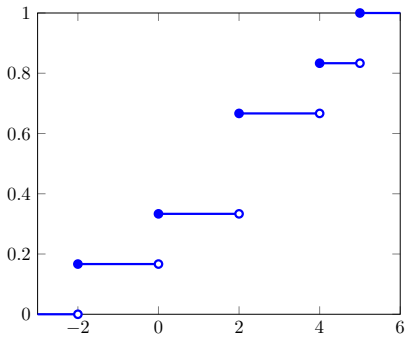
□

When dealing with cdfs, $F_X(a+)$ is used as an abbreviation of $\lim_{x\to a^+} F_X(x)$, and similarly, $F_X(a-)$ is used to abbreviate $\lim_{x\to a^-} F_X(x)$. Using this notation we can express the probabilities many events succinctly in terms of the cdf.

Event	Probability
$X \le a$	$F_X(a)$
$X < a$	$F_X(a-)$
$X = a$	$F_X(a) - F_X(a-)$
$X \ge a$	$1 - F_X(a-)$
$X > a$	$1 - F_X(a)$
$a < X < b$	$F_X(b-) - F_X(a)$
$a \le X \le b$	$F_X(b) - F_X(a-)$

Of particular interest is that fact that $P(X = a) = F_X(a) - F_X(a-)$. This is a formal statement of the observation that the possible values of a discrete random variable X are the values where F_X has a discontinuity, and the distance the graph jumps at a is $P(X = a)$.

Example 3.9 *Consider the function whose graph is given below. Check that this function a cdf for some random variable. If we call the random variable with this cdf Z, find $P(1 < Z < 2)$, $P(Z = 4)$, and $P(Z > 1)$.*



It's easy to see that this function is non-negative, non-decreasing, and tends to 0 as $x \to -\infty$ and 1 as $x \to \infty$ (we're assuming the graph continues off to the left and right along the horizontal lines shown). Furthermore, it's right-continuous since at each jump the value of the function is equal to the limit from the right.

$$P(1 < Z < 2) = F_Z(2-) - F_Z(1) = \frac{2}{6} - \frac{2}{6} = 0$$

$$P(Z = 4) = F_Z(4) - F_Z(4-) = \frac{5}{6} - \frac{4}{6} = \frac{1}{6}$$

$$P(Z > -1) = 1 - F_Z(-1) = 1 - \frac{1}{6} = \frac{5}{6}$$

3.2 Expected Value

Suppose we flip a fair coin three times and count the number of heads that appear. If we repeat this experiment many times, what do we expect the average number of heads to be?

If we let X be the number of heads observed when a single trial of this experiment is performed, then X is a discrete random variable whose distribution is given in the table below.

x	$P(X = x)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

If we perform the experiment many times, then we should expect to observe no heads about 12.5% of the time, one head 37.5% of the time, and so on (this is the frequentist interpretation of probability). If we weight each possible number of heads by the probability it occurs, we obtain

$$0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = 1.5.$$

So not surprisingly we should, on average, obtain 1.5 heads in every three flips of a fair coin. This notion of the average value of a random variable is formalized in the definition below.

Definition 3.5 *Given a discrete random variable X which takes values in the set $A = \{x_1, x_2, x_3, ...\}$, the expected value of X, also called the mean of X, is denoted $E(X)$ or μ_X, and is given by multiplying each possible value of X by its probability.*

$$\boxed{E(X) = \sum_{x \in A} x \cdot P(X = x) = \sum_{x \in A} x \cdot f_X(x)}$$

Remark The sum in the definition above could be infinite, in which case it may or not converge. If the sum is not absolutely convergent, then $E(X)$ is undefined.

Example 3.10 *Alice needs to decide whether to buy individual metro tickets at \$3.25 each, or an unlimited weekend pass at \$13.75. If she knows she'll take the metro either three, four, or five times this weekend with equal probability, what is the best decision?*

Let X represent the amount of money Alice will spend if she buys individual tickets. Then X is equally likely to take the three values 9.75, 13, *and* 16.25. *Therefore,*

$$E(X) = 9.75 \cdot \frac{1}{3} + 13 \cdot \frac{1}{3} + 16.25 \cdot \frac{1}{3} = 13.$$

On the other hand, if she purchases the unlimited pass, she'll pay \$13.75*. Thus, if Alice is interested in saving money, she should opt for the individual tickets.*

Example 3.11 *Consider the random variable Z whose cdf is given. Find $E(Z)$.*

$$F_Z(z) = \begin{cases} 0 & \text{if } z < -3 \\ \frac{1}{3} & \text{if } -3 \leq z < 1 \\ \frac{1}{2} & \text{if } 1 \leq z < 5 \\ \frac{5}{6} & \text{if } 5 \leq z < 6 \\ 1 & \text{if } z \geq 6 \end{cases}$$

Since $P(Z = a) = F_Z(a) - F_Z(a-)$ we can recover the pmf f_Z by examining the discontinuities of F_Z. From the pmf we compute $E(Z)$.

$$f_Z(z) = \begin{cases} \frac{1}{3} & \text{if } z = -3 \\ \frac{1}{6} & \text{if } z = 1 \\ \frac{1}{3} & \text{if } z = 5 \\ \frac{1}{6} & \text{if } z = 6 \end{cases}$$

$$E(Z) = \sum_{z \in A} z \cdot f_Z(z) = -3 \cdot \frac{1}{3} + 1 \cdot \frac{1}{6} + 5 \cdot \frac{1}{3} + 6 \cdot \frac{1}{6} = \frac{11}{6}$$

One immediate application of expected value is to games of chance. Given such a game, if we let X represent the player's profit from playing the game once, then if $E(X) > 0$ it's a winning bet, if $E(X) < 0$ it's a losing bet, and if $E(X) = 0$, it's a fair game.

Example 3.12 *On a roulette wheel, there are 18 black numbers, 18 red numbers, and 2 green numbers. Any wager made on black will double if successful. Find the expected value of a \$20 bet on black, assuming each of the numbers on the roulette wheel is equally likely to be selected.*

Let X represent the profit from a bet of \$20 *on black. Then X can take only values in $A = \{20, -20\}$ since the bet is either won or lost. There are a total of 38 possible outcomes, 18 of which result in a win, so $P(X = 20) = \frac{18}{38}$ and $P(X = -20) = \frac{20}{38}$.*

$$E(X) = 20 \cdot \frac{18}{38} + (-20) \cdot \frac{20}{38} = -\frac{40}{38} \simeq -1.05$$

Therefore, this is a losing bet. If you were to sit at a roulette table and repeatedly make \$20 bets on black for a very long time, you should expect your losses to total about \$1 per game played. Conversely for the casino, they should expect to make a \$1 profit for every \$20 bet on black played.

Example 3.13 *(St-Petersburg lottery) A coin is flipped until the first tail appears. You are initially given \$1 before the coin is flipped for the first time, and every time a head appears, the prize doubles. When the first tail appears, you receive the prize. If the sequence of flips that appears is $HHHT$, for example, you would receive \$8. What is the expected value of the prize?*

If we define a random variable X on the sample space $\Omega = \{T, HT, HHT, ...\}$ whose value is the prize, then X can take any value in $A = \{1, 2, 4, 8, 16, ...\}$, i.e., any value of the form 2^k for $k \in \mathbb{N}$. Furthermore, $X = 2^k$ when the outcome is a sequence with k heads followed by a tail. Thus, $P(X = 2^k) = \left(\frac{1}{2}\right)^{k+1}$.

$$\begin{aligned} E(X) &= \sum_{x \in A} xP(X = x) \\ &= \sum_{k=0}^{\infty} 2^k P(X = 2^k) \\ &= 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} + 8 \cdot \frac{1}{16} + ... \\ &= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + ... \to \infty \end{aligned}$$

The expected value is undefined, since the sum diverges by growing arbitrarily large.

This is a very strange result, since it implies that even if you're required to pay \$1 000 000 to play this game, it's still a winning bet. How can this be?

Notice that this is not a game one could actually play, as there's only a finite amount of money in the world. If this example is redone with an upper limit on the payout, the expected value will exist and give the average payout in the game. Using the current global GDP as the payout limit, the result is around \$45.

It's also worth considering that someone with $\$\,2 \times 10^{100}$ has twice as much money as someone with $\$\,1 \times 10^{100}$ in a literal sense, but both could spend the rest of their lives throwing briefcases of money into a fire and neither would make a dent in their net worth. In practice, twice as much money doesn't have twice as much value. Regardless, if one were actually able to play this game as stated, and play it as many times as one desires, it would be rational do so even if every play cost \$1 000 000, since one would *eventually* win an amount so large that the sum of all prior losses would be rendered insignificant.

Properties of the Expected Value

Theorem 3.2 *(Law of the Unconscious Statistician) If X is a discrete random variable that takes values in the set $A = \{x_1, x_2, x_3, ...\}$ and $g : \mathbb{R} \to \mathbb{R}$, the random variable $Y = g(X)$ has*

$$\boxed{E(Y) = \sum_{x \in A} g(x) f_X(x).}$$

Notice that the pmf in the sum is the pmf for X. We're finding the expected value of Y so we would imagine the pmf of Y would be involved here, but it isn't, and hence the name of the theorem. Using this result, we can calculate the expected value of Y without knowing its pmf, which saves a lot of work.

Pf. Consider $g(A) = \{g(x_1), g(x_2), g(x_3), ...\}$, the set of possible values for Y. From the definition of expected value, we have

$$E(Y) = \sum_{y \in g(A)} y \cdot f_Y(y).$$

Let $g^{-1}(y) = \{x \,|\, g(x) = y\}$, the set of all $x \in A$ that are sent to y by the function g. By definition, $f_Y(y)$ is equal to $P(Y = y)$, which is the sum of the probabilities of all values of X which g sends to y (we saw this in Example 3.7). Formally,

$$f_Y(y) = P(Y = y) = P(g(X) = y) = P(X \in g^{-1}(y)) = \sum_{x \in g^{-1}(y)} P(X = x).$$

Therefore, we have

$$\begin{aligned} E(Y) &= \sum_{y \in g(A)} y \cdot f_Y(y) = \sum_{y \in g(A)} y \cdot \sum_{x \in g^{-1}(y)} P(X = x) \\ &= \sum_{y \in g(A)} \sum_{x \in g^{-1}(y)} y \cdot P(X = x) = \sum_{y \in g(A)} \sum_{x \in g^{-1}(y)} g(x) \cdot P(X = x). \end{aligned}$$

In the last expression, the first sum is over all $y \in g(A)$, and for each such y, the second sum is over all $x \in A$ which are sent to that y. But every $x \in A$ is sent to some $y \in g(A)$, so the double sum is just summing over all $x \in A$. Thus,

$$E(Y) = \sum_{y \in g(A)} \sum_{x \in g^{-1}(y)} g(x) \cdot P(X = x) = \sum_{x \in A} g(x) \cdot P(X = x) = \sum_{x \in A} g(x) \cdot f_X(x).$$

□

Example 3.14 *Suppose that X is a random variable that takes the values -2, -1, 0, 1, and 2 with equal likelihood. Find the expected value of $Y = \frac{1}{1+X^2}$.*

Note that the pmf f_X has $f_X(x) = \frac{1}{5}$ for $x \in \{-2, -1, 0, 1, 2\}$, and that $Y = g(X)$, where $g(X) = \frac{1}{1+X^2}$. Using the Law of the Unconscious Statistician,

$$\begin{aligned} E(Y) &= \sum_{x \in A} g(x) f_X(x) \\ &= \frac{1}{1+(-2)^2} \cdot \frac{1}{5} + \frac{1}{1+(-1)^2} \cdot \frac{1}{5} + \frac{1}{1+(0)^2} \cdot \frac{1}{5} + \frac{1}{1+(1)^2} \cdot \frac{1}{5} + \frac{1}{1+(2)^2} \cdot \frac{1}{5} \\ &= \frac{1}{25} + \frac{1}{10} + \frac{1}{5} + \frac{1}{10} + \frac{1}{25} = \frac{12}{25} \end{aligned}$$

Corollary 3.1 *(Linearity of Expectation, Pt. I) If X is a discrete random variable, and $Y = aX + b$ for some $a, b \in \mathbb{R}$, then $E(Y) = aE(X) + b$.*

Pf. Applying the Law of the Unconscious Statistician, expanding, and then splitting up the sum, we have

$$\begin{aligned} E(Y) &= \sum_{x\in A} g(x)f_X(x) = \sum_{x\in A}(ax+b)\cdot f_X(x) = \sum_{x\in A} ax\cdot f_X(x) + bf_X(x) \\ &= \sum_{x\in A} ax\cdot f_X(x) + b\sum_{x\in A} f_X(x) = a\sum_{x\in A} x\cdot f_X(x) + b\sum_{x\in A} f_X(x) \\ &= aE(X) + b(1) = aE(X) + b. \end{aligned}$$

Note that $\sum_{x\in A} f_X(x) = 1$ since f_X is a probability mass function. □

With a clever application of this result, we can resolve the problem posed at the beginning of this chapter, which is a variation of the St-Petersburg lottery.

Example 3.15 *In a certain game, a coin is flipped until the first tail appears. If n heads occurred before this first tail which ends the game, the payout is $\$n$. What is the expected value of the payout?*

Let X be a random variable defined over the sample space $\Omega = \{T, HT, HHT, ...\}$ which gives the value of the payout. Then we have $P(X=0) = \frac{1}{2}$, $P(X=1) = \frac{1}{4}$, $P(X=2) = \frac{1}{8}$, and in general $P(X=k) = \frac{1}{2^k}$. Thus, $f_X(k) = \frac{1}{2^{k+1}}$ for $k \in \mathbb{N}$, and hence

$$E(X) = \sum_{x\in A} xf_X(x) = \sum_{k=1}^{\infty} k\cdot\frac{1}{2^{k+1}} = \frac{1}{4} + \frac{2}{8} + \frac{3}{16} + \frac{4}{32} + \frac{5}{64} + ...$$

In the St-Petersburg lottery, the infinite sum was clearly divergent, but this time it's not so easy to evaluate, in fact it's quite difficult.

Let's try another approach. Suppose the game has just begun and the first flip occurs. If the result is tails, then the game is over and you gain nothing, but if the result is heads, your payout goes up by $\$1$. In this second case, it's as if you've put $\$1$ into your pocket and started the game again, since each subsequent head also pays the same amount, $\$1$. If your payout in a new game is given by X, then your payout in a game in which a single head has already been flipped is given by $X+1$.

On the first flip then, there's a $\frac{1}{2}$ chance that a tail is flipped, and the payout is zero, and there's a $\frac{1}{2}$ chance that a head is flipped, and the payout is $X+1$. Therefore,

$$\begin{aligned} E(X) &= \tfrac{1}{2}\cdot 0 + \tfrac{1}{2}\cdot E(X+1) \\ E(X) &= 0 + \tfrac{1}{2}\cdot(E(X)+1) \\ E(X) &= \tfrac{1}{2}E(X) + \tfrac{1}{2} \\ E(X) - \tfrac{1}{2}E(X) &= \tfrac{1}{2} \\ E(X) &= 1 \end{aligned}$$

Corollary 3.1 was used to pass from the first to the second line, and from there on we simply isolated $E(X)$.

Notice that we can also obtain the expected value of any discrete random variable by summing over all possible outcomes $\omega \in \Omega$, instead of all possible values of X. This approach is sometimes more natural and sometimes more difficult.

Consider for example a random variable X that counts the number of heads in two flips of a fair coin. Using Definition 3.5, and letting $A = \{0, 1, 2\}$ denote the possible values of X, we have

$$E(X) = \sum_{x \in A} x f_X(x) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

On the other hand, we could let $X(\omega)$ denote the value of X when the outcome ω occurs, and sum over all outcomes in the sample space $\Omega = \{HH, HT, TH, TT\}$.

$$\begin{aligned} E(X) = \sum_{\omega \in \Omega} X(\omega)P(\omega) &= X(HH) \cdot \frac{1}{4} + X(HT) \cdot \frac{1}{4} + X(TH) \cdot \frac{1}{4} + X(TT) \cdot \frac{1}{4} \\ &= 2 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{4} = 1. \end{aligned}$$

Proposition 3.3 *For a discrete random variable X defined on the sample space Ω,*

$$E(X) = \sum_{\omega \in \Omega} X(\omega)P(\omega)$$

Theorem 3.3 *(Linearity of Expectation, Pt. II) If X and Y are discrete random variables defined over the same sample space Ω, then $E(X + Y) = E(X) + E(Y)$.*

Pf. Using the formula above to compute the expected value of $X + Y$, we have

$$\begin{aligned} E(X + Y) &= \sum_{\omega \in \Omega} (X(\omega) + Y(\omega))P(\omega) \\ &= \sum_{\omega \in \Omega} X(\omega)P(\omega) + Y(\omega)P(\omega) \\ &= \sum_{\omega \in \Omega} X(\omega)P(\omega) + \sum_{\omega \in \Omega} Y(\omega)P(\omega) \\ &= E(X) + E(Y). \end{aligned}$$

□

This result also extends to sums of any number of random variables defined on the same sample space. In general, the expected value of a sum of random variables is the sum of their expectations.

Example 3.16 *If we roll six dice and sum the results, what is the expected value of the sum?*

Let X_i denote the result of the i^{th} roll, and let $Y = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$. Every X_i has the same expected value, which we calculate below.

$$E(X_i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5$$

Thus, $E(Y) = 3.5 + 3.5 + 3.5 + 3.5 + 3.5 + 3.5 = 21$.

Example 3.17 *Suppose that Bob draws four marbles, with replacement, from an urn containing one blue, one green, one yellow, and one red marble. What is the expected number of different colours that will appear among Bob's selections?*

We'll solve this problem using a clever idea known as the method of indicators. Let I_b be an indicator variable for the event 'at least one blue marble is drawn', so I_b takes the value 1 if a blue marble is drawn and 0 if not. Define I_g, I_y, and I_r similarly, so we have an indicator variable for each colour.

The expected value of I_b is easy to calculate, since $I_b = 0$ when no blue marbles are drawn, so $P(I_b = 0) = (\frac{3}{4})^4 = \frac{81}{256}$, and hence $P(I_b = 1) = 1 - \frac{81}{256} = \frac{175}{256}$. Therefore, $E(I_b) = 0 \cdot \frac{81}{256} + 1 \cdot \frac{175}{256} = \frac{175}{256}$. The same is true for I_g, I_y, and I_r.

If we define X as the number of different colours that appear among the four draws, then $X = I_b + I_g + I_y + I_r$, hence by linearity of expectation,

$$\begin{aligned} E(X) &= E(I_b + I_g + I_y + I_r) \\ &= E(I_b) + E(I_g) + E(I_y) + E(I_r) \\ &= 4\left(\frac{175}{256}\right) = \frac{700}{256} \simeq 2.73. \end{aligned}$$

Notice that in the example above, though I_b, I_y, I_g, and I_r are indicators for events that are not independent, linearity of expectation allows us to compute the expected value of their sum without thinking about how the value of one affects any of the others.

Now consider a discrete random variable X whose range is $\{1, 2, \ldots, n\}$ for some positive integer n, and notice that if we define an indicator I_k for the event $X \geq k$ for each k from 1 to n, then $X = I_1 + I_2 + \ldots + I_n$.

To illustrate, say for example that X takes the value 2. Then I_1 and I_2 both take the value one, and all I_k for $k > 2$ are zero, so $X = I_1 + I_2 + I_3 + \ldots + I_n = 1 + 1 + 0 + \ldots + 0 = 2$. Using linearity of expectation, we have

$$E(X) = E(I_1 + I_2 + \ldots + I_n) = E(I_1) + E(I_2) + \ldots + E(I_n).$$

What is $E(I_k)$ for each $k \in \{1, 2, \ldots, n\}$? By definition, $I_k = 1$ if $X \geq k$ and $I_k = 0$ if $X < k$, hence $E(I_k) = 1 \cdot P(X \geq k) + 0 \cdot P(X < k) = P(X \geq k)$. Thus, $E(X) = E(I_1) + E(I_2) + \ldots + E(I_n) = P(X \geq 1) + P(X \geq 2) + \ldots + P(X \geq n)$.

Proposition 3.4 *(Tail Sum Formula) Given a discrete random variable X which takes values in $\{1, 2, \ldots, n\}$,*

$$\boxed{E(X) = \sum_{k=1}^{n} P(X \geq k).}$$

In fact, the assumption that there is an upper bound, n, on the values X can take turns out to be unnecessary, and we can sum over all positive integers.

Typically we compute expected values directly using Definition 3.5, but there are some cases where the method of indicators or this tail sum formula make calculations substantially faster.

Example 3.18 *For fair dice are rolled. What is the expected value of the smallest of the four resulting numbers?*

Let X be a random variable representing the smallest of the four resulting numbers. Then X takes values in the set $\{1, 2, 3, 4, 5, 6\}$, but writing the distribution of X involves doing a lot of combinatorics.

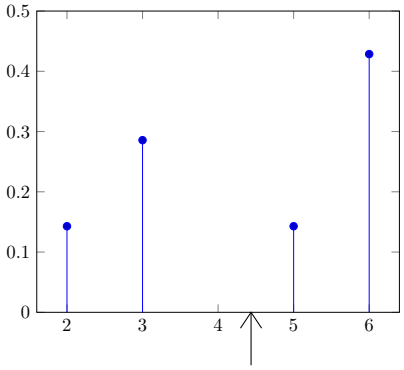
On the other hand, finding $P(X \geq k)$ for each $k \in \{1, 2, 3, 4, 5, 6\}$ is quite easy. For example, the event $X \geq 2$ occurs when every die shows a result higher than or equal to 2. Since the die rolls are independent, we have $P(X \geq 2) = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} = (\frac{5}{6})^4$. Using the tail sum formula yields

$$\begin{aligned} E(X) &= \sum_{k=1}^{6} P(X \geq k) \\ &= P(X \geq 1) + P(X \geq 2) + ... + P(X \geq 6) \\ &= 1 + \left(\frac{5}{6}\right)^4 + \left(\frac{4}{6}\right)^4 + \left(\frac{3}{6}\right)^4 + \left(\frac{2}{6}\right)^4 + \left(\frac{1}{6}\right)^4 \\ &\approx 1.76 \end{aligned}$$

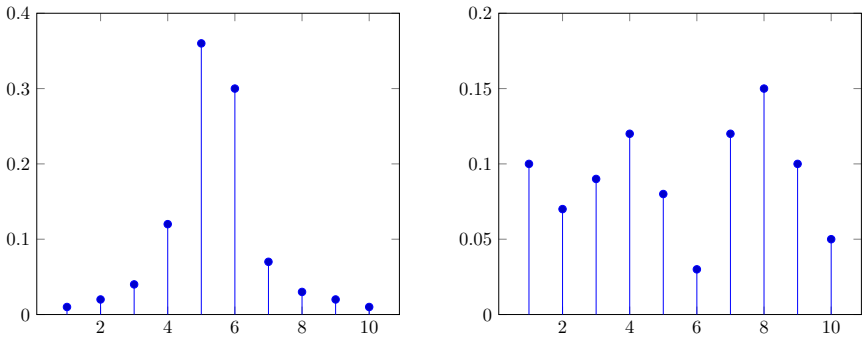
3.3 Variance

The expected value of a random variable gives us a way quantifying what's meant by the center of a distribution.

In fact, if the vertical lines in the graph of the pmf f_X each represent literal masses whose weights are proportional to their heights, then the value of $E(X)$ is the point on the x-axis where the distribution balances, and that's certainly a reasonable definition for the center.



We can also use expected value to quantify the amount of dispersion present in a probability distribution. Consider for example the two random variables X and Y whose pmfs are given on the left and right respectively.



The values of X tend to be clustered close to the center of the distribution, only very rarely will straying under 4 or over 7. On the other hand, Y takes on values far from the center of its distribution quite often. In other words, the distance between the value of X and the center of its distribution is usually smaller than the distance between the value of Y and its center of its distribution.

Remark In the discussions in this section, many expressions are made clearer by using the notation μ_X in place of $E(X)$, so keep in mind that μ_X is the expected value of X, which is a constant value at the center of the distribution of X.

Thus, we can quantify the dispersion of a random variable X by calculating the average distance from X to μ_X, that is, the expected value of the difference between X and its expected value. In practice, it's more common to work with the expected value of the *squared* difference between X and its average.

Definition 3.6 *The variance of a discrete random variable X is denoted $Var(X)$ or σ_X^2, and defined by $Var(X) = E[(X - \mu_X)^2]$. The standard deviation of X is then defined by $\sigma_X = \sqrt{\sigma_X^2} = \sqrt{Var(X)}$.*

Why square the differences between X and μ_X? It ensures that any deviations, either left or right of μ_X, are always counted positively. If we opt not to square, and ensure the differences are positive using the absolute value, we could calculate $E[\,|X - \mu_X|\,]$, which is called the mean absolute deviation.

Why variance and standard deviation? Why not use the mean absolute deviation defined above, or some other method to measure dispersion? We'll see later on that in fact, the variance measures exactly how difficult a random variable is to predict, in a mathematically precise sense. The ultimate goal in statistics is often to build predictive models, so for this reason variance is a practical measure. The standard deviation also plays a key role in many of the statistical testing procedures we'll see later in the course.

Example 3.19 *Determine the expected value, variance, and standard deviation of the random variable X whose pmf is given below.*

$$f_X(x) = \begin{cases} \frac{2}{8} & \text{if } x = 1 \\ \frac{1}{8} & \text{if } x = 3 \\ \frac{3}{8} & \text{if } x = 5 \\ \frac{2}{8} & \text{if } x = 6 \\ 0 & \text{otherwise} \end{cases}$$

First, $E(X) = 1 \cdot \frac{2}{8} + 3 \cdot \frac{1}{8} + 5 \cdot \frac{3}{8} + 6 \cdot \frac{2}{8} = 4$, that is, $\mu_X = 4$. Then we can calculate the variance below.

$$\begin{aligned} Var(X) &= E[(X - \mu_x)^2] \\ &= (1-4)^2 \cdot \frac{2}{8} + (3-4)^2 \cdot \frac{1}{8} + (5-4)^2 \cdot \frac{3}{8} + (6-4)^2 \cdot \frac{2}{8} \\ &= 9 \cdot \frac{2}{8} + 1 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 4 \cdot \frac{2}{8} = \frac{31}{8} \end{aligned}$$

Thus, the standard deviation of X is $\sigma_X = \sqrt{\frac{31}{8}} \simeq 1.968$.

Example 3.20 *In Big Win Bob's game, each player pays \$1 to place a bet on a number between 1 and 100, a winning number is selected at random, and the prize for a correct bet is \$75. In Conservative Charles' game, each player pays \$1 to bet on a number between 1 and 4, a winning number is selected at random, and the prize for a correct bet is \$3. Which game would you rather play?*

Let X represent a player's net gain in a single play of Bob's game, and let Y represent a player's net gain in a single play of Charles' game.

$$\begin{aligned} E(X) &= 74 \cdot \frac{1}{100} + (-1) \cdot \frac{99}{100} = -0.25 \\ E(Y) &= 2 \cdot \frac{1}{4} + (-1) \cdot \frac{3}{4} = -0.25 \end{aligned}$$

Thus, it does not matter which game you play in the long run. Over a very large number of plays the games are equivalent, either way the player loses 25¢ per game on average. However, the variances of X and Y are not equal.

$$\begin{aligned} Var(X) &= (74 - (-0.25))^2 \cdot \frac{1}{100} + (-1 - (-0.25))^2 \cdot \frac{99}{100} \simeq 55.69 \\ Var(Y) &= (2 - (-0.25))^2 \cdot \frac{1}{4} + (-1 - (-0.25))^2 \cdot \frac{3}{4} \simeq 1.68 \end{aligned}$$

This large difference in the variances indicates Big Win Bob's game is much riskier. In Bob's game players win rarely but win big, while in Charles' game players win small amounts quite often.

Calculating the variance by hand is a tedious process, and in practice, one would essentially always use a computer. In simple examples like those above, the following formula can make the calculation significantly faster.

Theorem 3.4 $\boxed{Var(X) = E[X^2] - (E[X])^2}$

Pf. Starting from the definition of $Var(X)$, expanding $(X-\mu_X)^2$ and using linearity of expectation, we have

$$\begin{aligned} Var(X) &= E[(X-\mu_X)^2] \\ &= E[X^2 - 2\mu_X \cdot X + {\mu_X}^2] \\ &= E[X^2] - E[2\mu_X \cdot X] + E[{\mu_X}^2]. \end{aligned}$$

Since μ_X is a constant, $E[{\mu_X}^2] = {\mu_X}^2$, and using linearity of expectation in the second term, $E[2\mu_X \cdot X] = 2\mu_X E[X]$. But $E[X]$ and μ_X are two different names for the same quantity, so $2\mu_X E[X] = 2\mu_X \cdot \mu_X = 2{\mu_X}^2$. Therefore,

$$\begin{aligned} Var(X) &= E[X^2] - E[2\mu_X \cdot X] + E[{\mu_X}^2] \\ &= E[X^2] - 2{\mu_X}^2 + {\mu_X}^2 \\ &= E[X^2] - {\mu_X}^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

. □

Example 3.21 *If X represents the outcome of a roll of a fair die, find σ_X.*

$$E[X] = \sum_{x=1}^{6} x f_X(x) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = \frac{7}{2}$$

$$E[X^2] = \sum_{x=1}^{6} x^2 f_X(x) = 1\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 9\cdot\frac{1}{6} + 16\cdot\frac{1}{6} + 25\cdot\frac{1}{6} + 36\cdot\frac{1}{6} = \frac{91}{6}$$

Hence, $Var(X) = E[X^2] - (E[X])^2 = \frac{91}{6} - (\frac{7}{2})^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}$, and square rooting, we obtain the standard deviation $\sigma_X \simeq 1.707$.

Properties of Variance

Proposition 3.5 *Let X be any discrete random variable with $Var(X) = 0$, then X takes a single value with probability one.*

Pf. Let A be the set of possible values of X, and suppose that $k_1, k_2 \in A$ such that $f_X(k_1) \neq 0$ and $f_X(k_2) \neq 0$, i.e. X can take either value with nonzero probability.

$$Var(X) = E[(X-\mu_X)^2] = \sum_{x\in A} (x-\mu_X)^2 f_X(x)$$

Every term in the sum on the right is non-negative since squares are non-negative, and $f_X(x) \geq 0$ by definition of a pmf. Furthermore, at least one term is nonzero

since k_1 and k_2 are distinct, so either $k_1 - \mu_X \neq 0$ or $k_2 - \mu_x \neq 0$. Thus, the sum must be positive, so we conclude that if X takes more than one distinct value with nonzero probability, $Var(X) \neq 0$. □

Theorem 3.5 *For any constants $a, b \in \mathbb{R}$,* $\boxed{Var(aX + b) = a^2 Var(X)}$.

Pf. This result follows from the definition of variance, using linearity of expectation.

$$\begin{aligned} Var(aX+b) &= E[((aX+b) - \mu_{aX+b})^2] \\ &= E[((aX+b) - E(aX+b))^2] \\ &= E[((aX+b) - (aE(X)+b))^2] \\ &= E[((aX - aE(X))^2] \\ &= E[a^2(X - E(X))^2] \\ &= a^2 E[(X - E(X))^2] \\ &= a^2 E[(X - \mu_X)^2] \\ &= a^2 Var(X) \end{aligned}$$

□

Example 3.22 *If X is a random variable with $E(X) = 2$ and $E[X(X-4)] = 7$, what is the standard deviation of $Y = 2X + 2$?*

Using linearity of expectation,

$$E[X(X-4)] = E[X^2 - 4X] = E[X^2] - 4E[X] = E[X^2] - 8.$$

Thus, $E[X^2] - 8 = 7$, so $E[X^2] = 15$, and hence

$$Var(X) = E[X^2] - (E[X])^2 = 15 - 4 = 11.$$

Applying the result above, $Var(Y) = Var(2X+2) = 4Var(X) = 44$, so $\sigma_Y = \sqrt{44}$.

Independent Random Variables

We've seen in Section 1.6 that two events $A \subseteq \Omega$ and $B \subseteq \Omega$ are independent if $P(A \mid B) = P(A)$, or equivalently, if $P(A \cap B) = P(A)P(B)$. This definition can be extended to discrete random variables in a natural way.

Definition 3.7 *Two random variables X and Y defined on the same sample space Ω are independent if $P(X = x \cap Y = y) = P(X = x)P(Y = y)$ for all x and y.*

Example 3.23 *Suppose a fair die is rolled twice. Let X be the result of the first roll and Y be the result of the second. Then X and Y are independent random variables, since for any x and y in $\{1, 2, 3, 4, 5, 6\}$, we have $P(X = x \cap Y = y) = \frac{1}{36}$, while $P(X = x) = \frac{1}{6}$ and $P(Y = y) = \frac{1}{6}$.*

Example 3.24 *Let X be a number chosen at random from the set $\{1,2,3,4,5,6\}$, and let Y be the remainder when X is divided by 3. Then X and Y are not independent random variables, since, for instance, $P(X=3 \cap Y=1)=0$, because if $X=3$ then we must have $Y=0$, while $P(X=3)P(Y=1)=\frac{1}{6}\cdot\frac{2}{6}$.*

In the last section we saw that $E(X+Y)=E(X)+E(Y)$ for any random variables X and Y. The relationship between $Var(X)$, $Var(Y)$, and $Var(X+Y)$ is generally more complicated. The details require defining the joint distribution of two random variables and an extensive discussion. If you're interested see [1] or [4]. In the case where X and Y are independent though, $Var(X+Y)=Var(X)+Var(Y)$. To prove this, we need to establish the lemma below.

Lemma 3.1 *If X and Y are independent random variables, $E(XY)=E(X)E(Y)$.*

Pf. Let A denote the range X and B denote the range of Y. To calculate $E(XY)$, we need to multiply every possible value of XY by the probability it occurs.

$$E(XY)=\sum_{(x,y)\,\in\, A\times B} xyP(X=x\cap Y=y)$$

Now $P(X=x\cap Y=y)=P(X=x)P(Y=y)=f_X(x)f_Y(y)$ since X and Y are independent. We can finish the argument by arranging the order of summation nicely and factoring.

$$\begin{aligned}E(XY)&=\sum_{(x,y)\,\in\, A\times B} xyf_X(x)f_Y(y)=\sum_{x\in A}\sum_{y\in B} xyf_X(x)f_Y(y)\\&=\sum_{x\in A} xf_X(x)\sum_{y\in B} yf_Y(y)=E(X)E(Y)\end{aligned}$$

□

Theorem 3.6 *If X and Y are independent, $Var(X+Y)=Var(X)+Var(Y)$.*

Pf. The result follows from a long calculation involving properties of expectation, and the lemma above, which is used in the second last line.

$$\begin{aligned}&Var(X+Y)=E[((X+Y)-E[X+Y])^2]\\&=E[(X+Y-E(X)-E(Y))^2]\\&=E[(X-E(X)+Y-E(Y))^2]\\&=E[(X-E(X))^2-2(X-E(X))(Y-E(Y))+(Y-E(Y))^2]\\&=E[(X-E(X))^2]-2E[(X-E(X))(Y-E(Y))]+E[(Y-E(Y))^2]\\&=Var(X)-2E[XY-XE(Y)-E(X)Y-E(X)E(Y)]+Var(Y)\\&=Var(X)-2(E[XY]-E[XE(Y)]-E[E(X)Y]-E[E(X)E(Y)])+Var(Y)\\&=Var(X)-2(E[X]E[Y]-E[X]E[Y]-E[X]E[Y]+E[X]E[Y])+Var(Y)\\&=Var(X)+Var(Y)\end{aligned}$$

□

Though we've only dealt with sums of two random variables, the arguments given above can be extended to finite sums of any length.

Corollary 3.2 *If X_1, X_2, ... , X_n is a sequence of independent random variables,*

$$\boxed{Var(X_1 + X_2 + \ ... \ + X_n) = Var(X_1) + Var(X_2) + \ ... \ + Var(X_n).}$$

Example 3.25 *Let X be the result of single roll of a die, and let Y be the average of the results of four rolls. Which random variable has the larger variance?*

We've seen in Example 3.21 that $Var(X) = \frac{35}{12}$. If we denote the results of four different die rolls by X_1, X_2, X_3, and X_4, then we have a sequence of independent random variables, and their average is $Y = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$.

$$\begin{aligned} Var(Y) &= Var(\tfrac{1}{4}(X_1 + X_2 + X_3 + X_4)) \\ &= \tfrac{1}{16}\, Var(X_1 + X_2 + X_3 + X_4) \\ &= \tfrac{1}{16}\, (Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4)) \\ &= \tfrac{1}{16}\, (\tfrac{35}{12} + \tfrac{35}{12} + \tfrac{35}{12} + \tfrac{35}{12}) \\ &= \tfrac{35}{48} \end{aligned}$$

The variance of the average of four rolls is four times smaller than the variance of a single roll. Note that this also means the standard deviation is two times smaller.

This example illustrates a very important principle that follows from the results of this section. If we take many measurements of the same quantity, each subject to some error, and the results of these measurements are independent, in the sense of Definition 3.7, then the result of averaging many of these measurements will give a more accurate estimate of the quantity being measured.

In fact, if we have a model for the error which affects each individual measurement, then we can use the results here to give very precise bounds on the number of measurements necessary to achieve a particular level of accuracy after averaging, and we'll do exactly this later in the course.

Moments

Expressions of the form $E[X^k]$ for some $k \in \mathbb{N}$ have been appearing quite frequently in our discussions in this section. Mathematicians call these values moments of the random variable X. There are a few common variations you might run into if you're reading about random variables, or statistics in general.

$$\begin{aligned} E[X^k] &\quad\longleftrightarrow\quad \text{The } k^{th} \text{ moment of } X. \\ E[(X - \mu_X)^k] &\quad\longleftrightarrow\quad \text{The } k^{th} \text{ central moment of } X. \\ E\left[\tfrac{1}{{\sigma_X}^k}(X - \mu_X)^k\right] &\quad\longleftrightarrow\quad \text{The } k^{th} \text{ standardized moment of } X. \end{aligned}$$

With this terminology, the second central moment is the variance of X. Many other moments have interesting names and applications, like skewness (the third standardized moment) and kurtosis (the fourth standardized moment).

3.4 The Uniform Distribution

Definition 3.8 *If X is a discrete random variable that takes a value in a finite set of consecutive integers $A = \{a, a+1, a+2, \ldots, a+n\}$, and X takes each value with equal probability, we say that X is uniformly distributed on A.*

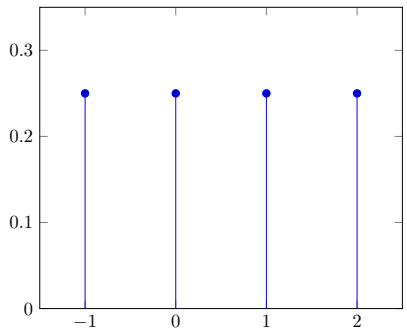
Example 3.26 *Suppose that X represents the roll of a single fair die, then X is uniformly distributed on $A = \{1, 2, 3, 4, 5, 6\}$, since $P(X = x) = \frac{1}{6}$ for all $x \in A$.*

The notation $X \sim Uniform(n)$ indicates X is a discrete random variable which is uniformly distributed on $A = \{1, 2, 3, .., n\}$. We could write $X \sim Uniform(6)$ to describe the distribution of the random variable in the example of a die roll above. However, note that for a general uniform distribution, A could be any set of consecutive integers, which need not begin with 1.

If X is uniformly distributed on the set A, then its pmf has the form

$$f_X(x) = \begin{cases} \frac{1}{n} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases},$$

which yields a graph with several bars of the same height. For example, the graph below is the pmf of a random variable uniformly distributed on $A = \{-1, 0, 1, 2\}$.



The expected value of a uniformly distributed variable on the set A is simply the average of the values in A. Formally, if X is uniformly distributed on the set $A = \{a_1, a_2, \ldots, a_n\}$, then

$$E(X) = a_1 \cdot \tfrac{1}{n} + a_2 \cdot \tfrac{1}{n} + \ldots + a_n \cdot \tfrac{1}{n} = \tfrac{a_1+a_2+\ldots+a_n}{n}.$$

In particular, if $X \sim Uniform(n)$, then $E(X) = \frac{1+2+3+\ldots+n}{n} = \frac{\frac{n(n+1)}{2}}{n} = \frac{n+1}{2}$.

Remark In general, if you wish to indicate that an item is being chosen from some set, and each item in that set has the same probability of being selected, you can say the item is chosen *uniformly at random* from the set. The term *uniformly* indicates that each item is equally likely.

3.5 The Bernoulli Distribution

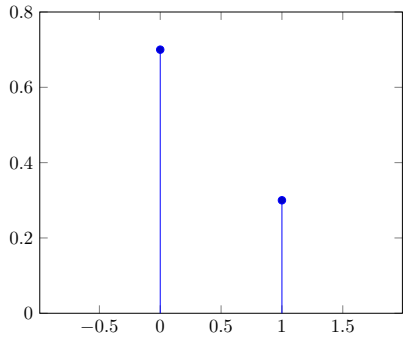
Suppose that we perform an experiment with only two possible outcomes, usually called success and failure, and the probability of success is known to be p. Such an experiment is called a Bernoulli trial, after Swiss mathematician Jacob Bernoulli. If X is a random variable that counts the number of successes in a single such trial, we say that X has a Bernoulli distribution.

Definition 3.9 *The random variable X has a Bernoulli distribution with parameter p, written $X \sim Bernoulli(p)$, if*

$$f_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}.$$

In other words, X takes the value 1 with probability p, and 0 with probability $1-p$.

Think of a Bernoulli random variable as counting the number of heads observed in a single flip of a coin with a given bias. If X represents the number of heads in a single flip of a fair coin, then $X \sim Bernoulli(0.5)$, and if Y represents the number of heads in a single flip of a biased coin which lands on heads only 30% of the time, then $Y \sim Bernoulli(0.3)$. The pmf of Y is shown below.



Theorem 3.7 *If $X \sim Bernoulli(p)$, then $E(X) = p$ and $Var(X) = p(1-p)$.*

Pf. These results both follow from direct calculations. From the probability mass function above we have $E(X) = 0 \cdot (1-p) + 1 \cdot p = p$. We can then calculate

$$\begin{aligned} Var(X) &= E[(X-\mu_X)^2] = E[(X-p)^2] = (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p \\ &= p^2(1-p) + (1-p)^2 p = p(1-p)(p + (1-p)) = p(1-p). \end{aligned}$$

□

The Bernoulli distribution is so simple that it's rarely useful for modelling any interesting phenomena (unless you'd call a single flip of a coin an interesting phenomenon). Its importance is that it allows us to define the Binomial, Geometric, and Poisson distributions, which all come up frequently in scientific practice.

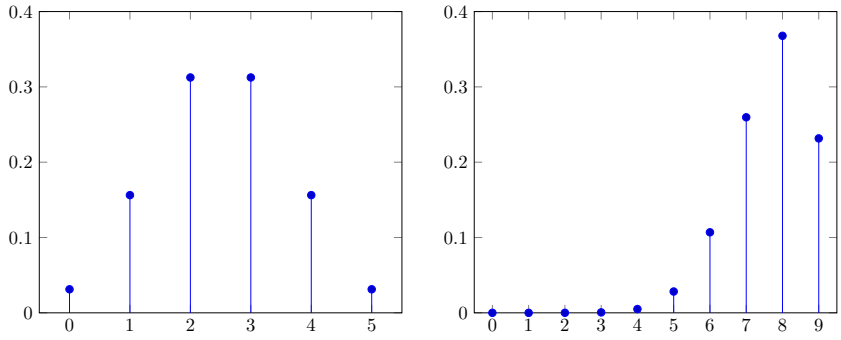
3.6 The Binomial Distribution

Suppose we perform a fixed number of Bernoulli trials (call this number n). Each trial has the same probability of success (call this value p), and the results of each trial form a sequence of independent Bernoulli random variables. If X counts the total number of successes after all the trials have been performed, then we say that X has a Binomial distribution.

Definition 3.10 *The random variable X has a binomial distribution with parameters n and p, written $X \sim Binomial(n, p)$, if*

$$f_X(x) = \begin{cases} \binom{n}{x}(p)^x(1-p)^{n-x} & \text{if } x = 0, 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}.$$

Below on the left is the pmf of a binomial distribution with $n = 5$, $p = 0.5$, and on the right, the pmf of a binomial distribution with $n = 9$, $p = 0.85$.



To derive this probability mass function, imagine recording of the results of the Bernoulli trials as an n letter long string of S's and F's, representing success and failure respectively. We saw in Section 2.2 on combinations that the number of such strings in which x successes occur is $\binom{n}{x}$.

Consider any such string, say $SFSSSFF$. In this case $n = 7$ and four successes occurred. The probability this sequence of outcomes occurs is not hard to calculate since we're performing independent Bernoulli trials, each with the same probability of success. Every S occurs with probability p, and every F with probability $1 - p$. Since there are four S's and three F's, and the trials are independent, this string occurs with probability $p^4(1-p)^3$.

In total then, there are $\binom{n}{x}$ sequences of outcomes with x successes, and each sequence occurs with probability $p^x(1-p)^{n-x}$, so $P(X = x) = \binom{n}{x}p^x(1-p)^{n-x}$.

The number of heads in n flips of a coin is the example that should immediately come to mind when you think of a binomial distribution, but binomial distributions also appear in many practical contexts involving samples.

Example 3.27 *Suppose that in a certain country's election,* 56% *of the voting population will vote for a particular candidate, Bob. If you select seven members of the voting population uniformly at random, and ask them who they'll vote for (assume they'll answer honestly), what is the probability your poll will incorrectly conclude that Bob will not win a majority of the vote?*

Let X represent the number of people who respond by saying that they'll vote for Bob. If we regard such a response as a success, and each interview is done without knowledge of the results of any of the others, then $X \sim Binomial(7, 0.56)$.

The result we're seeking is the probability that there are fewer than four successes, which we calculate from the pmf as follows.

$$\begin{aligned} P(X \leq 3) &= \sum_{i=0}^{3} P(X = i) = \sum_{i=0}^{3} \binom{7}{i} (0.56)^i (0.44)^{7-i} \\ &\simeq 0.0032 + 0.0284 + 0.1086 + 0.2304 \\ &= 0.3706 \simeq 37\% \end{aligned}$$

There is a 37% *chance that the results of this poll will incorrectly predict that Bob will not win a majority of the vote.*

Remark It would be easier to use the cdf to evaluate $P(X \leq 3)$, but unfortunately, the cdf for a binomial distribution does not have a simple closed form, so we're stuck summing the values of the pmf.

One of the reasons the binomial distribution appears in many contexts is that the process of evaluating a claim by doing repeated independent trials, which could be interviews, simulations, or literal laboratory experiments, is such a fundamental practice.

When we get into statistics later in the course, we'll show that as our sample of results grows larger, the probability of an incorrect prediction shrinks, and shrinks at a fast enough rate that it's possible to draw a reasonably accurate conclusion without having to use an unrealistically large sample.

Theorem 3.8 *If $X \sim Binomial(n, p)$, then $E(X) = np$ and $Var(X) = np(1-p)$.*

Pf. If X is binomial with parameters n and p, then X is the number of successes in n independent Bernoulli trials, which means $X = X_1 + X_2 + ... + X_n$, where each $X_i \sim Bernoulli(p)$. The expected value of each X_i is p by Theorem 3.7, so by linearity of expectation,

$$\begin{aligned} E(X) &= E(X_1 + X_2 + ... + X_n) \\ &= E(X_1) + E(X_2) + ... + E(X_n) \\ &= p + p + ... + p = np \end{aligned}$$

The variance can be derived similarly with the aid of Corollary 3.2.

$$\begin{aligned} Var(X) &= Var(X_1 + X_2 + ... + X_n) \\ &= Var(X_1) + Var(X_2) + ... + Var(X_n) \\ &= p(1-p) + p(1-p) + ... + p(1-p) = np(1-p) \end{aligned}$$

□

Example 3.28 *Let X be the number of doubles in five rolls of a pair of dice, and let Y be the number of heads in three flips of a fair coin. Which quantity is larger on average? Which quantity is more variable?*

X is a binomial random variable with $n = 5$ and $p = \frac{6}{36} = \frac{1}{6}$, so $E(X) = np = \frac{5}{6}$ and $Var(X) = np(1-p) = 5(\frac{1}{6})(\frac{5}{6}) = 25/36 \simeq 0.694$.

Y is a binomial random variable with $n = 3$ and $p = \frac{1}{2}$, so $E(Y) = np = \frac{3}{2}$ and $Var(X) = np(1-p) = 3(\frac{1}{2})(\frac{1}{2}) = 3/4 \simeq 0.75$.

Therefore, Y is larger on average. The second question is somewhat ambiguous, but if we use the variance as our measure, Y is a little more variable then X.

Example 3.29 *Edward lives in a country with very unreliable mail system. Only 7% of parcels sent in the mail make it to their destination. He wants to send two $20 books to his brother. If he sends them in one parcel, the postage is $5.20, and if he sends them separately, the postage is $3.30 for each book. To minimize the expected value of his expenses from a single attempt (loss + postage), which method is preferable?*

Let $X \sim Bernoulli(0.07)$, then sending both books in one parcel, the $5.20 postage charge always applies, and $40 is lost if $X = 1$. If L represents Edward's loss, then $L = 5.20 + 40X$, and therefore $E(L) = 5.20 + 40E(X) = 5.20 + 40 \cdot 0.07 = 8$.

On the other hand, if the two books are sent separately, then $L = 6.60 + 20Y$, where $Y \sim Binomial(2, 0.07)$. Thus, $E(L) = 6.60 + 20E(Y) = 6.60 + 20 \cdot 2 \cdot 0.07 = 9.4$, so sending both books in one package is the better option.

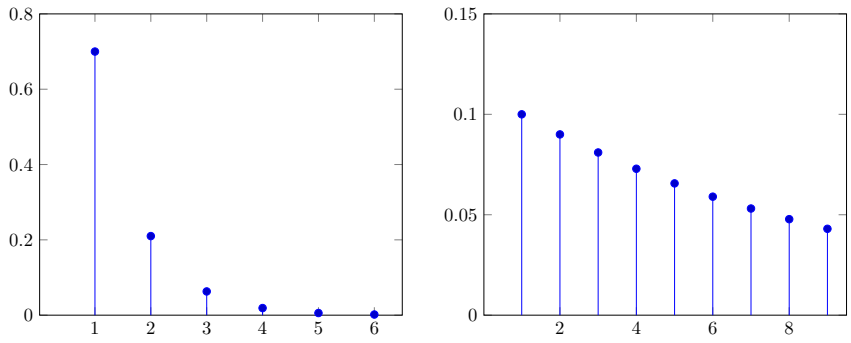
3.7 The Geometric Distribution

Binomial random variables result from performing independent Bernoulli trials a fixed number of times, and counting the number of successes. If instead, we repeat independent Bernoulli trials only until the first success occurs, and count the number of trials that were performed, we obtain a geometric random variable.

Definition 3.11 *The random variable X has a geometric distribution with parameter p, written $X \sim Geometric(p)$, if*

$$f_X(x) = \begin{cases} (1-p)^{x-1}p & \text{if } x = 1, 2, 3, ... \\ 0 & \text{otherwise} \end{cases}.$$

On the left is the pmf of a geometric distribution with $p = 0.7$, and on the right is the pmf of a geometric distribution with $p = 0.1$. Note that geometric random variables take arbitrarily large values, so both graphs continue to the right indefinitely, each approaching zero.



The derivation of the probability mass function can be done using the same ideas we used in the case of the binomial distribution. We can represent the outcomes of a sequence of repeated Bernoulli trials using S's for success and F's for failure. If we stop performing trials when the first success occurs, then the sequence of outcomes must be some number of F's followed by an S.

If the result was $FFFFFS$, then six trials were performed. Each failure occurs with probability $1 - p$, each success with probability p, and the trials are independent, so the probability this sequence occurs is $(1 - p)^5 p$. The same argument generalizes to any sequence of F's followed by an S and hence the probability that x trials must be performed before observing the first success is $(1 - p)^{x-1} p$.

Example 3.30 *Cards are repeatedly drawn with replacement from a shuffled deck. What is the probability that none of the first three cards that appear are face cards?*

If we define a success as drawing a face card, then the number of draws X until the first face card appears has $X \sim Geometric(\frac{12}{52})$.

The probability that none of the first three cards is a face card is $P(X > 3)$, which we can calculate from the pmf.

$$P(X > 3) = 1 - \sum_{x=1}^{3} P(X = x) = 1 - \frac{12}{52} - \frac{40}{52} \cdot \frac{12}{52} - \frac{40}{52} \cdot \frac{40}{52} \cdot \frac{12}{52} \simeq 0.455$$

In this example we could have made efficient use of the cdf to quickly calculate this probability. Let's write down the cdf for a geometric random variable so we can do this in the future. The proof is left as an exercise.

Proposition 3.6 *If $X \sim Geometric(p)$, the cdf of X is given by the expression*

below, where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x.

$$F_X(x) = \begin{cases} 1-(1-p)^{\lfloor x \rfloor} & \text{if } x \geq 1 \\ 0 & \text{if } x < 1 \end{cases}$$

In Example 3.30 above, $X \sim Geometric(\frac{12}{52})$, so $F_X(x) = 1-(1-\frac{12}{52})^x$ for all integers $x > 1$. We could obtain the result in that example via

$$P(X > 3) = 1 - P(X \leq 3) = 1 - F_X(3) = 1 - \left(1 - \left(1 - \left(\tfrac{12}{52}\right)\right)^3\right) \simeq 0.455.$$

Theorem 3.9 *If $X \sim Geometric(p)$, then $E(X) = \frac{1}{p}$ and $Var(X) = \frac{1-p}{p^2}$.*

Pf. The expected value can be calculated with clever use of derivatives. From the definition of expected value and the pmf of X,

$$E(X) = \sum_{x=1}^{\infty} x f_X(x) = \sum_{x=1}^{\infty} x(1-p)^{x-1}p = p\sum_{x=1}^{\infty} x(1-p)^{x-1}.$$

Observe that $\frac{d}{dp}(1-p)^x = x(1-p)^{x-1}(-1)$ and hence $\frac{d}{dp}(1-p)^x(-1) = x(1-p)^{x-1}$, regarding p as a variable and x as a constant. We can therefore rewrite the sum as

$$\begin{aligned} & p\sum_{x=1}^{\infty} \frac{d}{dp}(1-p)^x(-1) = -p\sum_{x=1}^{\infty} \frac{d}{dp}(1-p)^x \\ & \stackrel{!}{=} -p\frac{d}{dp}\left(\sum_{x=1}^{\infty}(1-p)^x\right) = -p\frac{d}{dp}\left(\frac{1-p}{1-(1-p)}\right) \\ & = -p\frac{d}{dp}\left(\frac{1-p}{p}\right). \end{aligned}$$

The infinite series was summed with the geometric series sum formula. To finish, we only need to evaluate the derivative using the quotient rule.

$$E(X) = -p\frac{d}{dp}\left(\frac{1-p}{p}\right) = -p\,\frac{(-1)(p)-(1-p)(1)}{p^2} = -p\frac{-1}{p^2} = \frac{1}{p}$$

□

We'll omit the argument for the variance. In fact, the same trick involving the derivative can be used twice to evaluate the infinite sum that defines $E[X^2]$, and we can obtain the variance with $Var(X) = E[X^2] - (E[X])^2$.

There is one step here that's suspect, the point where we factor the derivative out of the sum (marked with !). This works for finite sums, beacuse the sum rule in calculus tells us $\frac{d}{dp}(f(p)+g(p)) = \frac{d}{dp}f(p) + \frac{d}{dp}g(p)$. Does this same principle apply for infinite sums? Sometimes. The subtleties of this issue are beyond the scope of this course, but this instance turns out to be one where the move is justified.

Example 3.31 *Consider a noisy communications channel. There is a* 71% *chance that a message sent over this channel will become garbled and be unintelligible on the receiving end. Suppose that a message is sent until it's received intact without having become garbled, and each attempt takes 2 minutes. What is the probability the message will be correctly received in eight minutes or fewer? What is the average amount of time required before the message is received intact?*

If X *represents the number of attempts required to receive the message intact, then* $X \sim Geometric(0.29)$*, since there is an* 29% *chance of a successful transmission, and* $Y = 2X$ *is the amount of time required to receive the message intact. We can then easily calculate* $P(Y \leq 8)$ *and* $E(Y)$ *from the results in this section.*

$$\begin{aligned} P(Y \leq 8) &= P(X \leq 4) & E(Y) &= E(2X) \\ &= 1 - (1 - 0.29)^4 & &= 2E(X) \\ &\simeq 1 - 0.254 = 0.746 & &= 2 \cdot \tfrac{1}{0.29} \simeq 6.90 \end{aligned}$$

Thus, there is a 74.6% *chance the message will be correctly received in eight minutes or less, and the average amount of time required is* 6.90 *minutes.*

Notice that the formula $E(X) = 1/p$ has a remarkably intuitive interpretation. If one in every five people in a certain group drives a red car, then how many people from this group would you have to meet, on average, to find one who drives a red car? Five, right? In fact, that's precisely correct, since if $X \sim Geometric(\frac{1}{5})$, then $E(X) = 1/\frac{1}{5} = 5$.

In the concluding sections of this chapter, we've studied the Uniform, Bernoulli, Binomial, and Geometric distributions. If you look through a few probability texts, you might encounter the Poisson, Hypergeometric, and Negative Binomial distributions as well, but we can't cover everything, so we'll leave these out of our discussion.

Chapter 4

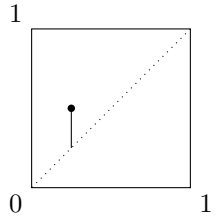
Continuous Random Variables

Introduction: Waiting For Dinner

Alice and Bob arrive home from work at independently determined random times between 5:00 pm and 6:00 pm. How long, on average, does the person who arrives earlier spend waiting for the one who arrives later?

If we measure Alice's arrival time in hours after 5:00 pm, then it can take any value in the interval $[0, 1]$. The same is true for Bob's arrival time, and the difference between the two arrival times. Random variables such as these which can take any value in an interval $I \subseteq \mathbb{R}$ are the subject of this chapter.

We can represent a pair of arrival times using a point in the unit square, with Alice's arrival time determining the horizontal coordinate, and Bob's determining the vertical coordinate. If Alice arrives first, the horizontal coordinate is smaller, so the point lies above the diagonal, and the time Alice must wait for Bob to arrive is the vertical distance to the diagonal (can you explain why?).



Similar remarks apply if Bob arrives first, in which case the time Bob must wait for Alice to arrive is the horizontal distance to the diagonal from the point representing the pair of arrival times.

Therefore, the solution to our problem is the average distance between a point in the unit square and the diagonal of the unit square, where we measure distance by moving only vertically or horizontally. At the moment, calculating this average is a daunting prospect. We need to develop the fundamentals of random variables that take values in intervals, and then we'll finish this problem in Section 4.4.

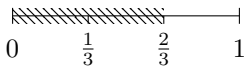
4.1 Continuous Probability Distributions

In Section 3.1, we defined a random variable as a function from Ω to $\mathbb{R}$, but in the chapter that followed, we only dealt with random variables that take values in some discrete subset $A \in \mathbb{R}$. We saw that we could specify the distribution of a random variable using either a table, a probability mass function, or a cumulative distribution function.

Our goal is now to understand what happens when a random variable can take any value in some interval $I \in \mathbb{R}$. Consider for example a random variable X which represents a number chosen at random from the interval $[0, 1]$.

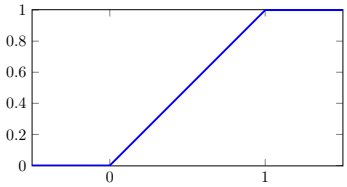
It's not possible to write the distribution of X as a table, since $[0, 1]$ is not a discrete set, that is, we can't write $[0, 1]$ as a list. Moreover, we saw in Section 1.3 that for any $x \in [0, 1]$, the probability X will take precisely that value is zero, so if we define the probability mass function of X as we did in Chapter 3, the resulting function is zero everywhere. The pmf is simply not a useful notion in this context.

Fortunately, the cumulative distribution function comes to the rescue. What is $P(X \leq \frac{2}{3})$? If we divide the interval into thirds, then $X \leq \frac{2}{3}$ when X falls into one of the left two thirds, which should happen two thirds of the time if X is equally likely to fall anywhere in $[0, 1]$.



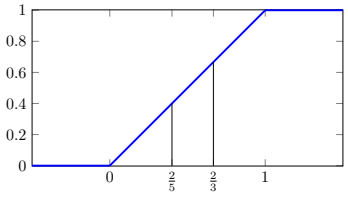
Thus, $P(X \leq \frac{2}{3}) = \frac{2}{3}$. In fact, by a similar argument, it should be clear that for any $x \in [0, 1]$, we have $P(X \leq x) = x$. Therefore, the cumulative distribution function of X is the function F_X given below.

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$



Using this cdf, we can determine the probability that X will be in various subsets of $[0, 1]$. Let's take the interval $(\frac{2}{5}, \frac{2}{3})$. Using the cdf above, we can compute

$$\begin{aligned} P(X \in (\tfrac{2}{5}, \tfrac{2}{3})) &= P(\tfrac{2}{5} < X < \tfrac{2}{3}) \\ &= P(X < \tfrac{2}{3}) - P(X \leq \tfrac{2}{5}) \\ &= F_X(\tfrac{2}{3}-) - F_X(\tfrac{2}{5}) \\ &= \tfrac{2}{3} - \tfrac{2}{5} = \tfrac{4}{15} \end{aligned}$$



Remark Recall $F_X(a-)$ is shorthand for $\lim_{x \to a^-} F_X(x)$, which gives $P(X < a)$. If the cdf of X is continuous, however, as is the case in our example above, then $\lim_{x \to a^-} F_X(x) = \lim_{x \to a^+} F_X(x) = F_X(a)$, and hence $F_X(a-)$, $F_X(a+)$ and $F_X(a)$ are all equal. When dealing with continuous cdfs we'll simply omit the $-$ or $+$ from now on since it makes no difference in our results.

Note that this implies a random variable X with a continuous cdf will take any fixed real number value with probability zero, since

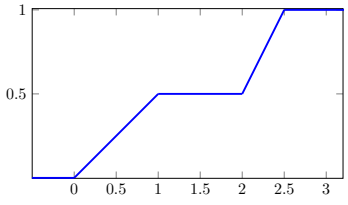
$$P(X = x) = P(X \leq x) - P(X < x) = F_X(x) - F_X(x-) = 0.$$

It makes sense to ask for the probability that such a random variable falls into a certain range or interval, but the probability it's exactly equal to a given value is always zero. This implies that $P(X \in (a, b))$, $P(X \in [a, b))$, $P(X \in (a, b])$, and $P(X \in [a, b])$ are all equal.

Definition 4.1 *If X is a random variable whose cumulative distribution function $F_X(x) = P(X \leq x)$ is continuous, then we say X is a continuous random variable.*

Example 4.1 *Consider the continuous random variable X whose cdf is given below. Find $P(X \in (\frac{1}{3}, 1) \cup [\frac{3}{2}, \frac{5}{2}])$.*

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2}x & \text{if } 0 \leq x < 1 \\ \frac{1}{2} & \text{if } 1 \leq x < 2 \\ x - \frac{3}{2} & \text{if } 2 \leq x < 3 \\ 1 & \text{if } x \geq 3 \end{cases}$$



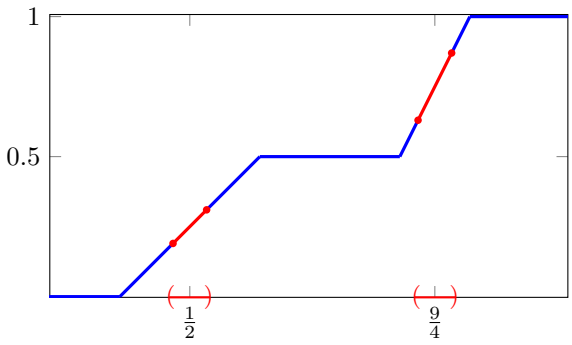
Since the events $X \in (\frac{1}{3}, 1)$ and $X \in [\frac{3}{2}, \frac{5}{2}]$ are mutually exclusive,

$$\begin{aligned} P(X \in (\tfrac{1}{3}, 1) \cup [\tfrac{3}{2}, \tfrac{5}{2}]) &= P(X \in (\tfrac{1}{3}, 1)) + P(X \in [\tfrac{3}{2}, \tfrac{5}{2}]) \\ &= P(X < 1) - P(X \leq \tfrac{1}{3}) + P(X \leq \tfrac{5}{2}) - P(X < \tfrac{3}{2}) \\ &= F_X(1) - F_X(\tfrac{1}{3}) + F_X(\tfrac{5}{2}) - F_X(\tfrac{3}{2}) \\ &= \tfrac{1}{2} - \tfrac{1}{2} \cdot \tfrac{1}{3} + (\tfrac{5}{2} - \tfrac{3}{2}) - \tfrac{1}{2} \\ &= \tfrac{1}{2} - \tfrac{1}{6} + 1 - \tfrac{1}{2} = \tfrac{5}{6}. \end{aligned}$$

Probability Density Functions

Consider a continuous random variable X, and two real numbers r and s. We now know that $P(X = r)$ and $P(X = s)$ are both zero, but we can still ask whether X is more likely to end up 'close to' r or 'close to' s.

We can formalize this by asking for the probability X lies in each of the tiny intervals $(s - \epsilon, s + \epsilon)$ and $(r - \epsilon, r + \epsilon)$, for some very small positive number ϵ. Take for instance the random variable X in Example 4.1 above. Is it more likely this variable takes a value close to $\frac{1}{2}$, or that it takes a value close to $\frac{9}{4}$?

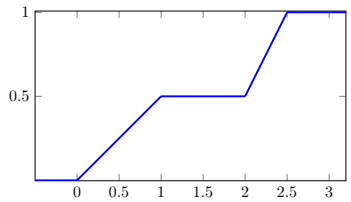


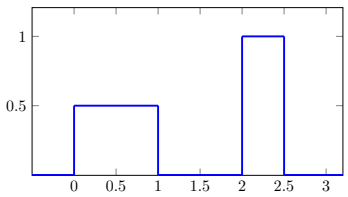
The probability X lies in the interval (a, b) is $F_X(b) - F_X(a)$, the difference in the value of the cdf at the endpoints. For the cdf shown above, the graph is twice as steep when it's near $\frac{9}{4}$ than it is when it's near $\frac{1}{2}$, and hence this difference is twice as large. Therefore, X is twice as likely to assume values near $\frac{9}{4}$ than it is to assume values near $\frac{1}{2}$.

In general then, the probability X lies close to a given value is proportional to the slope of the cdf at that value, but we know how to measure the slope of any function, with the derivative. Thus, given any $x \in \mathbb{R}$, if we evaluate the derivative of F_X at x, the larger the result, the more likely it is that X takes a value very close to x. Remember that the probability $X = x$ must be zero for all $x \in \mathbb{R}$, so when we evaluate the derivative of F_X at x, the result is not a probability, instead it's known as a probability density, and the derivative of F_X is called the probability density function of X.

Definition 4.2 *If X is a continuous random variable with cdf F_X, the probability density function of X is denoted f_X and is defined by $f_X(x) = \frac{d}{dx} F_X(x)$.*

The cdf F_X of the continuous random variable X defined in Example 4.1 above is shown once again on the left, and the corresponding probability density function f_X is on the right. It's standard practice to draw vertical lines on the graph of the probability density function as below. These lines are technically not part of the graph, but can make it easier to read.





We'll often use pdf to abbreviate probability density function. The rationale for the term 'probability density' is the following: when considering a rod of metal or other material, we can only ask for the mass of a contiguous piece of the material, since the mass at a point is always zero. We can, however, consider the density of the material to be well defined at a point, as the rate of change in mass at that point as we move along the rod. The situation with continuous random variables is analogous: a probability density measures the rate of change of the cdf as we move along it, in other words, the rate at which probability is accumulating as we move past that point.

Proposition 4.1 *If X is a continuous random variable, then the probability X lies in the interval (a, b) is the area under f_X on (a, b).*

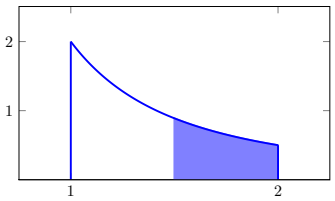
Pf. Calculating $P(X \in (a, b))$ using the cdf F_X as we have done in the examples above, and applying the fundamental theorem of calculus, we have

$$P(X \in (a, b)) = F_X(b) - F_X(a) = \int_a^b \frac{d}{dx} F_X(x)\, dx = \int_a^b f_X(x)\, dx.$$

This integral is the area under f_X on (a, b). Note that the result would be the same with any of the intervals $(a, b]$, $[a, b)$, and $[a, b]$. □

Example 4.2 *Let X be the continuous random variable whose pdf is given below. Find $P(X > \frac{3}{2})$.*

$$f_X(x) = \begin{cases} \frac{2}{x^2} & \text{if } 1 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$



$$P(X > \tfrac{3}{2}) = \int_{\frac{3}{2}}^{2} f_X(x)\, dx = \int_{\frac{3}{2}}^{2} \frac{2}{x^2}\, dx = -\frac{2}{x}\Big|_{\frac{3}{2}}^{2} = -\frac{2}{2} - \left(-\frac{2}{\frac{3}{2}}\right) = \frac{1}{3}$$

Proposition 4.2 *If X is any continuous random variable with cdf F_X, the pdf f_X satisfies each of the properties below.*

(i) For all $x \in \mathbb{R}$, $f_X(x) \ge 0$.

(ii) $\int_{-\infty}^{\infty} f_X(x)\,dx = P(-\infty < X < \infty) = 1.$

(iii) $\int_{-\infty}^{t} f_X(x)\,dx = P(X \leq t) = F_X(t).$

Intuitively, any non-negative function which encloses an area of one between its graph and the horizontal axis is the pdf of some random variable whose cdf can be recovered through integration, as in *(iii)* above.

Remark In fact, the technical details here get difficult and subtle. The mathematical world is teeming with exotic functions which behave in extremely strange ways, but regardless, this intuition is a good one to have.

Example 4.3 *Consider the function f given below. Find c so that this function is the pdf of some continuous random variable.*

$$f(x) = \begin{cases} cxe^{-x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

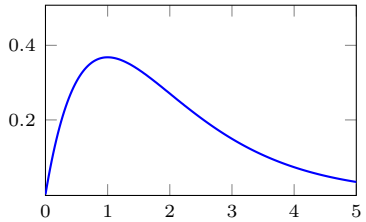
Since $e^x > 0$ always, we have $xe^{-x} > 0$ when $x > 0$, and thus f is non-negative when $c > 0$. We must find a positive value for c such that the area under f is one, so we evaluate the area under f.

$$\int_{\infty}^{\infty} f(x)\,dx = \int_{0}^{\infty} cxe^{-x}\,dx = c\int_{0}^{\infty} xe^{-x}\,dx = c(-x-1)e^{-x}\Big|_{0}^{\infty}$$

The last equality comes from using integration by parts. To complete the evaluation, we can use L'Hôpital's rule.

$$\begin{aligned} c(-x-1)e^{-x}\Big|_{0}^{\infty} &= \lim_{x\to\infty}\left(c(-x-1)e^{-x}\right) - c(-0-1)e^{0} \\ &= c\lim_{x\to\infty}\left(\tfrac{-x-1}{e^x}\right) + c \\ &\overset{H}{=} c\lim_{x\to\infty}\left(\tfrac{-1}{e^x}\right) + c \\ &= 0 + c = c \end{aligned}$$

The area under f is simply c, and therefore if we take $c = 1$, f is a pdf. The graph of f with $c = 1$ is given below.



Functions of a Random Variable

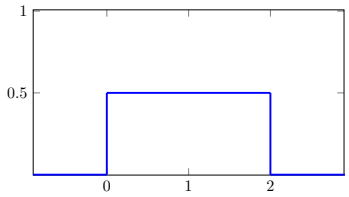
If X is any random variable and $g : \mathbb{R} \to \mathbb{R}$ is a function, then as we've seen in the discrete case, we can define a new random variable $Y = g(X)$. Given the pdf f_X, how can we obtain f_Y? Two methods are demonstrated in the examples below, the method of distribution functions and the method of substitution.

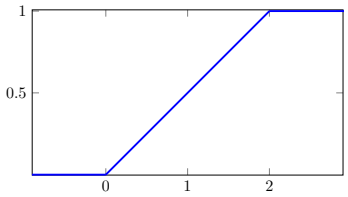
Example 4.4 *Let X be a real number chosen at random in the interval $[0, 2]$, that is, $f_X(x) = \frac{1}{2}$ if $x \in [0, 2]$, and f_X is zero otherwise. Find the pdf of $Y = X^2 + 1$ by the method of distribution functions.*

It's not hard to see that the function F_X below is an antiderivative of f_X which goes from zero to one as it progresses along the x axis, and hence is the cdf of X.

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2}x & \text{if } 0 \le x \le 2 \\ 1 & \text{if } x > 2 \end{cases}$$

The functions f_X and F_X are shown below on the left and right respectively.





Now that we have the cdf of X, we can use it to determine the cdf of Y as below.

$$F_Y(y) = P(Y \le y) = P(X^2 + 1 \le y) = P(X \le \sqrt{y-1}) = F_X(\sqrt{y-1})$$

Hence if we replace x by $\sqrt{y-1}$ in the cdf of X, we obtain the cdf of Y. Note that the range of the function $g(x) = x^2 + 1$ on the interval $[0, 2]$ is $[1, 5]$, and hence the cdf of Y is given below.

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 1 \\ \frac{1}{2}\sqrt{y-1} & \text{if } 1 \le y \le 5 \\ 1 & \text{if } y > 5 \end{cases}$$

We now obtain $f_Y(y) = \frac{1}{2} \cdot \frac{1}{2\sqrt{y-1}} = \frac{1}{4\sqrt{y-1}}$ on $[1, 5]$ by differentiating the cdf F_Y.

The method of distribution functions is very natural, but its disadvantage is that it requires knowledge of the cdf of X. In the example above, it was not difficult to determine F_X from f_X, but in general, finding an antiderivative won't always be so easy. We can avoid this issue with the method of substitution.

Example 4.5 *Let X be a random variable with pdf $f_X(x) = \frac{2}{x^2}$ on $[1,2]$. Find the pdf of $Y = \ln(X)$ by the method of substitution.*

The method of substitution works just like integration by substitution, so it should look familiar. We'll begin with the relationship $y = \ln(x)$, and differentiate to obtain a relationship between the differentials dy and dx.

$$y = \ln(x) \quad \to \quad dy = \frac{1}{x}\,dx \quad \to \quad dx = x dy \quad \to \quad dx = e^y dy$$

Next, we rewrite $f_X(x)\,dx$ in terms of the variable y and the differential dy.

$$f_X(x)\,dx = \frac{2}{x^2}\,dx = \frac{2}{(e^y)^2}\,e^y dy = \frac{2}{e^y}\,dy = f_Y(y)\,dy$$

Therefore, $f_Y(y) = \frac{2}{e^y}$ on $[0, \ln(2)]$. As with the last example, we can determine the set of values where Y is nonzero by taking the range of the function $g(x) = \ln(x)$ on the interval $[1,2]$.

This method works whenever $g : \mathbb{R} \to \mathbb{R}$ is a differentiable one-to-one function. Why? It follows from integration by substitution. Consider random variables X and $Y = g(X)$, and note that $X \in [a,b]$ exactly when $Y \in [g(a), g(b)]$ for any $a, b \in \mathbb{R}$, so

$$\int_a^b f_X(x)\,dx = P(X \in [a,b]) = P(Y \in [g(a), g(b)]) = \int_{g(a)}^{g(b)} f_Y(y)\,dy.$$

Substituting $y = g(x)$ in the integral on the left will indeed change the limits of integration from a and b to $g(a)$ and $g(b)$, as desired, and the value of the integral will not change either (if it did, integration by substitution wouldn't be a correct way to evaluate integrals), so the equality will hold.

What needs to be done then is to simply perform the substitution $y = g(x)$ and rewrite the integrand $f_X(x)\,dx$ in terms of the new variable y. This is exactly what we did in the example above. Whatever function appears afterwards in front of the differential dy will be the pdf $f_Y(y)$ of our random variable Y.

Remark There's a detail above we need to be careful about. We assumed g was an increasing function when we wrote $Y \in [g(a), g(b)]$, since we need $g(a) < g(b)$ for this interval to make sense. Any continuous one-to-one function must be either strictly increasing or strictly decreasing. If g is decreasing, we'll have $g(a) > g(b)$, so the interval for Y we want is in fact $[g(b), g(a)]$. However, using properties of integrals we can write

$$\int_{g(a)}^{g(b)} f_Y(y)\,dy = -\int_{g(b)}^{g(a)} f_Y(y)\,dy = \int_{g(b)}^{g(a)} -f_Y(y)\,dy = P(Y \in [g(b), g(a)])$$

and so if it turns out g is decreasing, our resulting pdf $f_Y(y)$ will be off by a sign (hence never positive!). In this case, we can just swap the sign on $f_Y(y)$.

4.2 Expected Value

The expected value of a continuous random variable X is defined in essentially the same way the expected value of a discrete random variable was defined in Section 3.2. In place of a probability mass function we have a probability density function, and in place of a sum we have the continuous analogue, an integral. We'll also see in this section that all the nice properties of expectation we derived in Section 3.1 still hold for continuous random variables.

Definition 4.3 *If X is a continuous random variable, the expected value or mean of X is denoted $E(X)$ or μ_X, and defined by the integral below.*

$$\boxed{E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx}$$

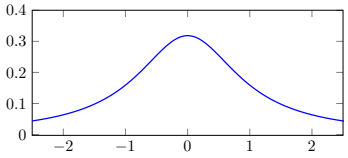
Note that this improper integral may not converge, in which case we say that the expected value is undefined.

Example 4.6 *Find the expected value of the random variable X whose pdf is given by $f_X(x) = 2/x^2$ on $[1, 2]$.*

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_1^2 x\, \frac{2}{x^2}\, dx = 2\int_1^2 \frac{1}{x}\, dx = 2\ln(x)\Big|_1^2 = 2\ln(2)$$

Example 4.7 *Consider the random variable X whose pdf is given below (called a Cauchy distribution). Show that $E(X)$ is undefined.*

$$f_X(x) = \frac{1}{\pi(1+x^2)} \quad \text{on} \quad (-\infty, \infty)$$



$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_{-\infty}^{\infty} x \frac{1}{\pi(1+x^2)}\, dx = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x}{1+x^2}\, dx$$

To evaluate a 'doubly improper' integral like this, we must split it into two improper integrals to see if both of those converge. In the two resulting integrals, we make the substitution $u = 1 + x^2$, so $du = 2x\, dx$ and hence $dx = \frac{1}{2x} du$.

$$\begin{aligned} E(X) &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x}{1+x^2}\, dx \\ &= \frac{1}{\pi}\int_{-\infty}^{0} \frac{x}{1+x^2}\, dx + \frac{1}{\pi}\int_{0}^{\infty} \frac{x}{1+x^2}\, dx \\ &= \frac{1}{\pi}\int_{\infty}^{1} \frac{x}{u}\frac{1}{2x}\, du + \frac{1}{\pi}\int_{1}^{\infty} \frac{x}{u}\frac{1}{2x}\, du \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2\pi}\int_{\infty}^{1} \frac{1}{u}\,du + \frac{1}{2\pi}\int_{1}^{\infty} \frac{1}{u}\,du \\
&= \frac{1}{2\pi}\ln(u)\Big|_{\infty}^{1} + \frac{1}{2\pi}\ln(u)\Big|_{1}^{\infty} \\
&= \frac{1}{2\pi}\left(\ln(1) - \lim_{u\to\infty} ln(u)\right) + \frac{1}{2\pi}\left(\lim_{u\to\infty} ln(u) - \ln(1)\right)
\end{aligned}
$$

Neither of the two improper integrals converge since $\lim_{u\to\infty}\ln(u) = \infty$*, and hence* $E(X)$ *is undefined.*

Theorem 4.1 *(Law of the Unconscious Statistician) If* X *is a continuous random variable and* $g : \mathbb{R} \to \mathbb{R}$ *then the expected value of* $Y = g(X)$ *is given below.*

$$\boxed{E(Y) = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx}$$

Example 4.8 *Suppose a 1 metre long stick is broken at a position chosen at random along its length. What is the average length of the longer piece?*

If X *denotes the position where the break occurs, then the pdf of* X *is* $f_X(x) = 1$ *on* $[0, 1]$*, and the two pieces will have lengths* x *and* $1 - x$*. The piece of length* x *is longer when* $x > \frac{1}{2}$ *and the piece of length* $1 - x$ *is longer when* $x < \frac{1}{2}$*. Thus,* $Y = g(X)$ *where*

$$g(x) = \begin{cases} 1 - x & \text{if } 0 \le x \le \frac{1}{2} \\ x & \text{if } \frac{1}{2} \le x \le 1 \end{cases}.$$

Now we can calculate the expected value of Y *by using the Law of the Unconscious Statistician.*

$$
\begin{aligned}
E(Y) &= \int_{-\infty}^{\infty} g(x) f_X(x)\,dx = \int_{0}^{1} g(x)\cdot 1\,dx \\
&= \int_{0}^{\frac{1}{2}} g(x)\,dx + \int_{\frac{1}{2}}^{1} g(x)\,dx \\
&= \int_{0}^{\frac{1}{2}} 1 - x\;dx + \int_{\frac{1}{2}}^{1} x\,dx \\
&= (x - \tfrac{1}{2}x^2)\Big|_{0}^{\frac{1}{2}} + \tfrac{1}{2}x^2\Big|_{\frac{1}{2}}^{1} \\
&= (\tfrac{1}{2} - \tfrac{1}{8}) - (0 - 0) + \tfrac{1}{2} - \tfrac{1}{8} = \tfrac{3}{4}
\end{aligned}
$$

Hence, the average length of the longer piece is $\frac{3}{4}$*.*

Proposition 4.3 *(Linearity of Expectation) If* X *and* Y *are continuous random variables, then* $E(aX + b) = aE(X) + b$ *and* $E(X + Y) = E(X) + E(Y)$*.*

The arguments for these linearity results mirror the corresponding arguments given for the discrete versions in Section 3.2, but everywhere we had a sum, we'll now have an integral. The first property above, for example, can be derived using the Law of the Unconscious Statistician as follows.

$$\begin{aligned} E(aX+b) &= \int_{-\infty}^{\infty} (ax+b) f_X(x)\, dx \\ &= \int_{-\infty}^{\infty} axf_X(x) + bf_X(x)\, dx \\ &= a\int_{-\infty}^{\infty} xf_X(x)\, dx + b\int_{-\infty}^{\infty} f_X(x)\, dx \\ &= aE(X) + b\cdot 1 = aE(X) + b \end{aligned}$$

Compare this to the discrete version, Corollary 3.1 in Section 3.2, and you'll see that it's exactly the same after changing each sum to an integral. Note that a double sum in the discrete case will become a double integral in the continuous one, and the topic of double integrals deserves a more thorough treatment than we have time for in this course. For this reason, we'll omit the argument for the law of the unconscious statistician above.

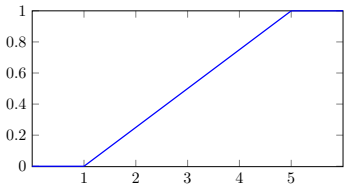
The continuous analogue of the tail sum formula (Proposition 3.4) can be derived from an argument involving double integrals as well.

Proposition 4.4 *(Continuous Tail Sum Formula) If X is a continuous random variable whose range contains only non-negative values, then*

$$\boxed{E(X) = \int_0^{\infty} P(X > x)\, dx = \int_0^{\infty} 1 - F_X(x)\, dx.}$$

Example 4.9 *Suppose that X is a randomly selected real number between 1 and 5. Then the pdf of X is $f_X(x) = \frac{1}{4}$ on $[1,5]$, and the cdf of X is given below.*

$$F_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{1}{4}(x-1) & \text{if } 1 \le x < 5 \\ 1 & \text{if } x \ge 5 \end{cases}$$



We can tell that $E(X) = \frac{1+5}{2} = 3$ by symmetry (recall the expected value is the literal balance point of the graph of the pdf). Verify that the tail sum formula yields the same result.

$$\begin{aligned}
E(X) &= \int_0^\infty 1 - F_X(x)\,dx \\
&= \int_0^1 1 - 0\ dx + \int_1^5 1 - \frac{1}{4}(x-1)\,dx + \int_5^\infty 1 - 1\ dx \\
&= \int_0^1 1\ dx + \int_1^5 \frac{5}{4} - \frac{1}{4}x\ dx \\
&= x\Big|_0^1 + \left(\frac{5}{4}x - \frac{1}{8}x^2\right)\Big|_1^5 \\
&= 1 + \left(\frac{25}{4} - \frac{25}{8}\right) - \left(\frac{5}{4} - \frac{1}{8}\right) \\
&= 1 + \frac{25}{8} - \frac{9}{8} = 3
\end{aligned}$$

This example may not convince you the continuous tail sum formula is very useful, but it should at least provide some reassurance that it's correct. We'll see examples where the tail sum formula will make our calculations easier in the next few sections.

4.3 Variance

Definition 4.4 *The variance of a continuous random variable X is denoted $Var(X)$ or ${\sigma_X}^2$, and is defined as follows.*

$$\boxed{Var(X) = E[(X-\mu_X)^2] = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\,dx}$$

The variance of X is undefined if this improper integral diverges.

As in the discrete case, the standard deviation of a continuous random variable is defined by $\sigma_X = \sqrt{Var(X)}$, and the two useful identities below hold.

Proposition 4.5 *If X is a continuous random variable and $Var(X)$ is defined,*

(i) $Var(X) = E[X^2] - (E[X])^2$

(ii) $Var(aX+b) = a^2 Var(X)$

Example 4.10 *The proportion of parts produced at a certain factory each day which pass quality control is a random variable with pdf $f_X(x) = 105x^4(1-x)^2$ on $[0,1]$. Find $E(X)$ and $Var(X)$.*

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^1 105x^5(1-x)^2\,dx = 105\int_0^1 x^5 - 2x^6 + x^7\,dx \\
&= 105\left(\frac{1}{6}x^6 - \frac{2}{7}x^7 + \frac{1}{8}x^8\right)\Big|_0^1 = 105\left(\frac{1}{6} - \frac{2}{7} + \frac{1}{8}\right) = 0.625
\end{aligned}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^1 105x^6(1-x)^2\,dx = 105\int_0^1 x^6 - 2x^7 + x^8\,dx$$
$$= 105\left(\frac{1}{7}x^7 - \frac{2}{8}x^8 + \frac{1}{9}x^9\right)\Big|_0^1 = 105\left(\frac{1}{7} - \frac{2}{8} + \frac{1}{9}\right) \simeq 0.4166$$

Thus, $E(X) = 0.625$ *and* $Var(X) = E[X^2] - (E[X])^2 \simeq 0.4166 - 0.3906 = 0.0260$.

The variance measures the dispersion of a random variable, and so allows to us to compare two random variables on the basis of how much they vary, but so far we haven't been able to make any concrete statements about the values a random variable can take based only on its expected value and variance. The result below allows us to do this by giving a probabilistic bound on how far a random variable can stray from its expected value.

Theorem 4.2 *(Chebyshev's Inequality) For any random variable* X *whose expected value and variance are defined, the probability that* X *is within* k *standard deviations of its expected value is at least* $1 - \frac{1}{k^2}$.

$$\boxed{P\big(X \in (\mu_X - k\sigma_X,\, \mu_X + k\sigma_X)\big) \geq 1 - \frac{1}{k^2}}$$

Pf. Consider splitting the integral in the definition of $Var(X)$ at the endpoints of the interval $(\mu_X - k\sigma_X,\, \mu_X + k\sigma_X)$.

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\,dx$$
$$= \int_{-\infty}^{\mu_X - k\sigma_X} (x-\mu_X)^2 f_X(x)\,dx + \int_{\mu_X - k\sigma_X}^{\mu_X + k\sigma_X} (x-\mu_X)^2 f_X(x)\,dx$$
$$+ \int_{\mu_X + k\sigma_X}^{\infty} (x-\mu_X)^2 f_X(x)\,dx$$

Note that for the last integral, $x \in (\mu_X + k\sigma_X, \infty)$, which means $x > \mu_X + k\sigma_X$, so $(x-\mu_X)^2 > k^2{\sigma_X}^2$. We can make a similar argument if $x \in (-\infty, \mu_X + k\sigma_X)$ and conclude $(x-\mu_X)^2 > k^2{\sigma_X}^2$ in that case also. Thus, if we replace $(x-\mu_X)^2$ with $k^2{\sigma_X}^2$ in the first and third integrals, and remove the second integral, whose value must be non-negative since f_X is a non-negative function, we can only have diminished the right hand side.

$$Var(X) \geq \int_{-\infty}^{\mu_X - k\sigma_X} k^2{\sigma_X}^2 f_X(x)\,dx + \int_{\mu_X + k\sigma_X}^{\infty} k^2{\sigma_X}^2 f_X(x)\,dx$$
$$\geq k^2{\sigma_X}^2\left(\int_{-\infty}^{\mu_X - k\sigma_X} f_X(x)\,dx + \int_{\mu_X + k\sigma_X}^{\infty} f_X(x)\,dx\right)$$
$$\geq k^2{\sigma_X}^2\left[P\left(X \in (-\infty, \mu_X - k\sigma_X)\right) + P\left(X \in (\mu_X + k\sigma_X, \infty)\right)\right]$$
$$\geq k^2{\sigma_X}^2\left[1 - P\left(X \in (\mu_X - k\sigma_X, \mu_X + k\sigma_X)\right)\right]$$

Since $Var(X) = {\sigma_X}^2$, dividing both sides by ${\sigma_X}^2$ (which is always positive) yields the inequality below.

$$\begin{aligned} 1 &\geq k^2 \left[1 - P\left(X \in (\mu_X - k\sigma_X,\, \mu_X + k\sigma_X)\right)\right] \\ \frac{1}{k^2} &\geq 1 - P\left(X \in (\mu_X - k\sigma_X,\, \mu_X + k\sigma_X)\right) \\ -1 + \frac{1}{k^2} &\geq -P\left(X \in (\mu_X - k\sigma_X,\, \mu_X + k\sigma_X)\right) \\ 1 - \frac{1}{k^2} &\leq P\left(X \in (\mu_X - k\sigma_X,\, \mu_X + k\sigma_X)\right) \end{aligned}$$

□

Example 4.11 *On a fishing trip, a fisherman catches ten fish and claims three of them weigh over* $10\,kg$*. If the average weight of the ten fish is* $5.2\,kg$*, and the standard deviation is* $2.4\,kg$*, show that the claim must be false.*

Let X *represent the weight of a randomly selected fish in the group of ten, and note that* $10 = 5.2 + 2(2.4)$*, so a weight of* $10\,kg$ *is exactly two standard deviations above the mean. We can use Chebyshev's inequality to determine the minimum number of fish whose weights are within two standard deviations of the mean.*

$$P\left(X \in (5.2 - 2(2.4),\, 5.2 + 2(2.4)\right) \geq 1 - \frac{1}{2^2} = \frac{3}{4}$$

Thus, at least 75% *of the fish must be between* $5.2 - 2(2.4) = 0.4\,kg$ *and* $5.2 + 2(2.4) = 10\,kg$*, which means that at most* 25% *can weigh over* $10\,kg$*. Three fish over* $10\,kg$ *would represent a proportion of* 30%*, contradicting Chebyshev's Inequality. Therefore, the claim is false.*

Remark Note that in Example 4.11 above, the random variable X, which represents the weight of a randomly selected fish from a certain group of ten, is discrete, since it can only have ten or fewer distinct values. Though the proof was given for the continuous case, Chebyshev's Inequality applies in the discrete case as well.

Chebyshev's Inequality plays an important role in the theoretical foundations of statistics because of its generality. The distribution of X (equivalently, the graph of its pdf f_X) might have any ridiculous shape you could imagine, and yet as long as $E(X)$ and $Var(X)$ are well defined, this inequality gives some simple probabilistic constraints on the value of X.

4.4 The Uniform Distribution

Definition 4.5 *If the probability density function of* X *is constant on some interval* $I \subseteq \mathbb{R}$*, and zero outside* I*, then* X *is said to be uniformly distributed on the interval* I*, written* $X \sim Uniform(I)$*.*

Remark Be careful not to confuse continuous and discrete uniform distributions. The notation $X \sim Uniform(5)$ describes a discrete random variable that takes values in the set $\{1,2,3,4,5\}$ while $X \sim Uniform([1,5))$ refers to a continuous random variable that takes values in the interval $[1,5)$.

Example 4.12 *Let X be uniformly distributed on $[3,7]$. Find f_X and F_X.*

To determine f_X, we're looking for a positive constant c such that the area under $f_X(x) = c$ on $[3,7]$ is one. This area is a rectangle with width 4 and height c, hence we require $4c = 1$, so $c = \frac{1}{4}$.

$$f_X(x) = \begin{cases} \frac{1}{4} & \text{if } 3 \le x \le 7 \\ 0 & \text{otherwise} \end{cases}$$

To find F_X, we have $F_X(x) = \int f_X(x)\,dx = \int \frac{1}{4}\,dx = \frac{1}{4}x + C$ on $[3,7]$, and since $F_X(7) = P(X \le 7) = 1$, we have $\frac{1}{4}(7) + C = 1$, so $C = -\frac{3}{4}$.

$$F_X(x) = \begin{cases} 0 & \text{if } x < 3 \\ \frac{1}{4}x - \frac{3}{4} & \text{if } 3 \le x \le 7 \\ 1 & \text{if } x > 7 \end{cases}$$

In general, if X is uniformly distributed on an interval I which has its left endpoint at a and its right endpoint at b, then $f_X(x) = \frac{1}{b-a}$ on I and f_X is zero everywhere else. Next, let's derive the expected value and variance of a continuous uniform random variable.

Proposition 4.6 *If I is an interval with endpoints a and b, and $X \sim Uniform(I)$, then $E(X) = \frac{a+b}{2}$ and $Var(X) = \frac{1}{12}(b-a)^2$.*

Pf. We observed above that $f_X(x) = \frac{1}{b-a}$ on I, so we calculate $E(X)$ and $E(X^2)$.

$$E(X) = \int_a^b x f_X(x)\,dx = \int_a^b x\,\tfrac{1}{b-a}\,dx = \tfrac{1}{b-a}\int_a^b x\,dx = \tfrac{1}{b-a}\left(\tfrac{1}{2}x^2\Big|_a^b\right) = \tfrac{a+b}{2}$$

$$E(X^2) = \int_a^b x^2 f_X(x)\,dx = \int_a^b x^2\,\tfrac{1}{b-a}\,dx = \tfrac{1}{b-a}\int_a^b x^2\,dx = \tfrac{1}{b-a}\left(\tfrac{1}{3}x^3\Big|_a^b\right) = \tfrac{a^2+ab+b^2}{3}$$

Note that the difference of squares and difference of cubes identities were used above to simplify the results.

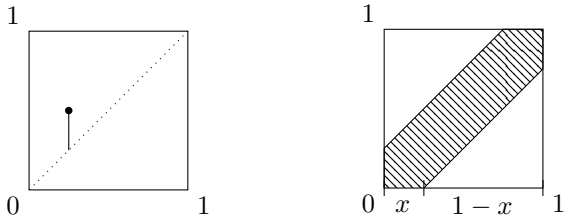
$$Var(X) = E[X^2] - (E[X])^2 = \tfrac{a^2+ab+b^2}{3} - \tfrac{a^2+2ab+b^2}{4} = \tfrac{a^2-2ab+b^2}{12} = \tfrac{1}{12}(b-a)^2$$

□

Though the uniform distribution is certainly the most basic type of continuous random variable we'll study, there are many interesting problems and applications that involve uniformly distributed quantities. The problem in the introduction to this chapter is a great example.

Example 4.13 *Alice and Bob each arrive home from work independently at times uniformly distributed between 5:00 pm and 6:00 pm every day. How long, on average, does the person who arrives earlier spend waiting for the one who arrives later?*

In the introduction, we noted that we can represent pairs of arrival times as points in the unit square, Alice's arrival time on the horizontal axis, and Bob's on the vertical axis, both in hours after 5:00 pm. The waiting time is given by the vertical distance to the diagonal for points above it, and the horizontal distance to the diagonal for points below it. Thus, points in the shaded region shown below on the right occur when Alice and Bob arrive within x hours of each other.



If both the x and y coordinates (the arrival times) are uniformly distributed on the axes, the point is uniformly distributed in the square, which has unit area. Thus, the probability that Alice and Bob arrive within x hours of each other is given by the shaded area above, which can be found by subtracting out the areas of the two triangles, to obtain $P(X \leq x) = 1 - 2(\frac{1}{2}(1-x)^2) = 1 - (1-x)^2$.

Thus, $P(X \leq x) = 1 - (1-x)^2$ if $x \in [0,1]$ and $P(X \leq x) = 1$ if $x > 1$. We can now calculate $E(X)$ using the tail sum formula.

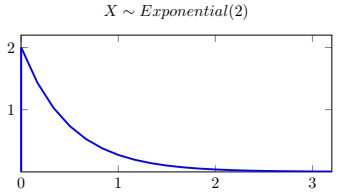
$$E(X) = \int_0^\infty P(X > x)\,dx = \int_0^1 1 - P(X \leq x)\,dx = \int_0^1 1 - (1-(1-x)^2)\,dx$$
$$= \int_0^1 (1-x)^2\,dx = \int_0^1 1 - 2x + x^2\,dx = x - x^2 + \tfrac{1}{3}x^3\Big|_0^1 = \tfrac{1}{3}$$

Thus, the expected waiting time is one third of an hour, or twenty minutes.

4.5 The Exponential Distribution

Definition 4.6 *The random variable X is exponentially distributed with parameter $\lambda > 0$ (often called the rate parameter), written $X \sim Exponential(\lambda)$, if its pdf is given by the function below.*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



Exponential random variables are used to model time between recurrent events in scenarios where observing such an event gives no information about when the next event will occur. These recurring events are often called arrivals, and the rate parameter λ gives the average rate at which the arrivals occur.

Quantities that are typically modelled by exponential random variables include times required for radioactive atoms to decay, distances between mutations on a DNA sequence, and lifetimes of mechanical or electronic devices.

Proposition 4.7 *If $X \sim Exponential(\lambda)$, then its cdf is the function given below, and $E(X) = \frac{1}{\lambda}$.*

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

Pf. To obtain the cdf F_X, let's find an antiderivative for f_X on $[0, \infty)$ using the substitution $u = -\lambda x$, so $du = -\lambda dx$, and $dx = \frac{-1}{\lambda} du$.

$$\begin{aligned} \int f_X(x)\, dx &= \int \lambda e^{-\lambda x}\, dx \\ &= \int \lambda e^u \frac{-1}{\lambda}\, du \quad \text{where} \quad u = -\lambda x \\ &= -\int e^u\, du \\ &= -e^u + C \\ &= -e^{-\lambda x} + C \end{aligned}$$

Thus, $F_X(x) = -e^{-\lambda x} + C$ on $[0, \infty)$, and since $F_X(0) = P(X \leq 0) = 0$, we have $e^0 + C = 0$, hence $C = 1$ and $F_X(x) = 1 - \lambda e^{-\lambda x}$ as desired. Now, using the tail sum formula,

$$\begin{aligned} E(X) &= \int_0^\infty 1 - F_X(x)\, dx \\ &= \int_0^\infty 1 - (1 - e^{-\lambda x})\, dx \\ &= \int_0^\infty e^{-\lambda x}\, dx \\ &= \int_0^{-\infty} e^u \frac{-1}{\lambda}\, du \quad \text{where} \quad u = -\lambda x \\ &= \frac{-1}{\lambda} \int_0^{-\infty} e^u\, du \\ &= \frac{-1}{\lambda} \left(\lim_{u \to -\infty} e^u - e^0 \right) \\ &= \frac{-1}{\lambda} (0 - 1) = \frac{1}{\lambda}. \end{aligned}$$

□

Example 4.14 *At a certain intersection, there is an average of three car accidents per day. Assuming the time between crashes is exponentially distributed, what is the probability that the next two accidents will occur at least five hours apart?*

Let X denote the gap between consecutive crashes in hours. Then X is exponentially distributed and $E(X) = \frac{1}{\lambda} = 8$ since there is an average of 3 crashes in 24 hours, so the average gap between crashes is 8 hours. Thus, $\lambda = \frac{1}{8}$. Note that we can also interpret this as a rate, one crash per eight hours on average.

$$\begin{aligned} P(X \in (5,\infty)) &= 1 - \int_0^5 f_X(x)\,dx \\ &= 1 - \int_0^5 \frac{1}{8} e^{-\frac{1}{8}x}\,dx \\ &= 1 - \int_0^{-\frac{5}{8}} \frac{1}{8} e^u (-8\,du) \quad \textit{where} \quad u = -\frac{1}{8}x \\ &= 1 + \int_0^{-\frac{5}{8}} e^u\,du \\ &= 1 + \left(e^u \big|_0^{-\frac{5}{8}} \right) \\ &= 1 + (e^{-\frac{5}{8}} - 1) \simeq 0.535 \end{aligned}$$

The variance of an exponential random variable can be derived from the formula $Var(X) = E[X^2] - (E[X])^2$ after computing $E[X^2]$ using integration by parts. The result is given below, and the proof is left as an exercise.

Proposition 4.8 *If $X \sim Exponential(\lambda)$, then $Var(X) = \frac{1}{\lambda^2}$.*

One key property of the exponential distribution is that it's 'memoryless'. If X is an exponentially distributed random variable representing, for example, a gap between recurring events, then regardless of how long you've already been waiting for the next event to occur, the probability that it will occur in the next minute is the same. Formally, we can state the memoryless property as follows.

Proposition 4.9 *(Memoryless Property) If X is an exponentially distributed random variable, and $a, b \in \mathbb{R}$, then $P(X > a + b \,|\, X > b) = P(X > a)$.*

Example 4.15 *Suppose that the lifetime of a certain model of monitor is exponentially distributed, with an average of ten years. If one of these monitors has already lasted fifteen years, what is the probability it will last at least five more years?*

To show the memoryless property holds, we'll calculate $P(X > 20 \,|\, X > 15)$, and then check that we would obtain the same result with $P(X > 5)$. Since the average lifetime is ten years, we have $E(X) = \frac{1}{\lambda} = 10$, hence $\lambda = \frac{1}{10}$.

$$\begin{aligned} P(X > 20) &= 1 - P(X \le 20) = 1 - (1 - e^{-\frac{1}{10}\cdot 20}) = e^{-2} \\ P(X > 15) &= 1 - P(X \le 15) = 1 - (1 - e^{-\frac{1}{10}\cdot 15}) = e^{-\frac{3}{2}} \end{aligned}$$

$$P(X > 20 \,|\, X > 15) = \frac{P(X > 20 \cap X > 15)}{P(X > 15)} = \frac{P(X > 20)}{P(X > 15)} = \frac{e^{-2}}{e^{-\frac{3}{2}}} = e^{-\frac{1}{2}}$$

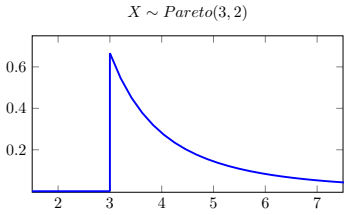
The probability the monitor will last at least five more years is $e^{-\frac{1}{2}} \simeq 0.607$*, and the memoryless property states that we'll obtain the same result by simply calculating*

$$P(X > 5) = 1 - P(X \leq 5) = 1 - (1 - e^{-\frac{1}{10} \cdot 5}) = e^{-\frac{1}{2}}.$$

4.6 The Pareto Distribution

Definition 4.7 *The random variable* X *is Pareto distributed with the two parameters* $m > 0$ *(the minimum) and* $c > 0$ *(often called the shape parameter), written as* $X \sim Pareto(m, c)$*, if its pdf is given by the function below.*

$$f_X(x) = \begin{cases} \frac{cm^c}{x^{c+1}} & \text{if } \ x \geq m \\ 0 & \text{otherwise} \end{cases}$$



This distribution is named for Italian engineer and economist Vilfredo Pareto, who noted that 80% of land in Italy was owned by 20% of the population. If we extend this idea further and imagine that of those 20% who own the most land, 80% of the land they own is held by the top 20% of them, and so on, then the distribution of the amount of land owned by a randomly selected Italian individual will be well approximated by a Pareto random variable. This '80-20 rule' or similar proportionality rules appear in many contexts, and when they do the Pareto distribution is typically involved.

Pareto distributions have been used to model household income, sizes of cities (by population), magnitudes of earthquakes, and sizes of insurance claims.

Example 4.16 *If the population of a randomly selected US city is approximated by a Pareto distributed random variable* $X \sim Pareto(40\,000, 2.3)$*, what proportion of cities have a population between 100 000 and 1 000 000?*

$$\begin{aligned} P(100\,000 < X < 1\,000\,000) &= \int_{100\,000}^{1\,000\,000} \frac{2.3 \cdot 40\,000^{\,2.3}}{x^{3.3}}\, dx \\ &= 2.3 \cdot 40\,000^{\,2.3} \int_{100\,000}^{1\,000\,000} \frac{1}{x^{3.3}}\, dx \\ &= 2.3 \cdot 40\,000^{\,2.3} (\tfrac{-1}{2.3} x^{-2.3}) \Big|_1^{100} \\ &= -40\,000^{\,2.3} (x^{-2.3}) \Big|_1^{100} \simeq 12.1\% \end{aligned}$$

Proposition 4.10 *If $X \sim Pareto(m, c)$, then $E(X) = \frac{cm}{c-1}$ if $c > 1$, and $E(X)$ is undefined if $c \leq 1$.*

Pf. Let $X \sim Pareto(m, c)$, then

$$E(X) = \int_{\infty}^{\infty} x f_X(x)\,dx = \int_m^{\infty} x\frac{cm^c}{x^{c+1}}\,dx = cm^c \int_m^{\infty} \frac{1}{x^c}\,dx$$
$$= cm^c \frac{-1}{(c-1)} x^{-(c-1)}\Big|_m^{\infty} = \frac{cm^c}{c-1}\left(-\left(\lim_{x\to\infty} \frac{1}{x^{c-1}}\right) + \frac{1}{m^{c-1}}\right).$$

Note that the argument above does not apply if $c = 1$, since when we integrate $\frac{1}{x}$, we would obtain $\ln(x)$. However, if $c \neq 1$, we obtain the expression above, and the value of the resulting limit depends only on whether the exponent on x is positive or negative.

$$\lim_{x\to\infty} \frac{1}{x^{c-1}} = \begin{cases} 0 & \text{if } c > 1 \\ \infty & \text{if } c < 1 \end{cases}$$

Thus, $E(X)$ is undefined if $c < 1$, and $E(X) = \frac{cm^c}{c-1}\left(\frac{1}{m^{c-1}}\right) = \frac{cm}{c-1}$ if $c > 1$. We need to deal with the case $c = 1$ separately.

$$E(X) = \int_{\infty}^{\infty} x f_X(x)\,dx = \int_m^{\infty} x\frac{1 \cdot m^1}{x^2}\,dx = m\int_m^{\infty} \frac{1}{x}\,dx$$
$$= m\ln(x)\Big|_m^{\infty} = m\left(\left(\lim_{x\to\infty} \ln(x)\right) - \ln(m)\right) \to \infty$$

□

We can determine $E(X^2)$ by a very similar calculation. Applying the formula $Var(X^2) = E[X^2] - (E[X])^2$ yields the result below.

Proposition 4.11 *If $X \sim Pareto(m, c)$, then $Var(X) = \frac{cm^2}{(c-1)^2(c-2)}$ if $c > 2$, and $Var(X)$ is undefined if $c \leq 2$.*

Example 4.17 *Suppose that a Pareto distribution is used to model a quantity with an average value of 4 and a variance of 2. What are the parameters m and c?*

Since $E(X) = 4$, we have $4 = \frac{cm}{c-1}$, so $c - 1 = \frac{cm}{4}$. Substituting this result into the variance formula given in Proposition 4.11 above,

$$2 = \frac{cm^2}{(c-1)^2(c-2)} = \frac{cm^2}{\frac{c^2m^2}{16}(c-2)} = \frac{16}{c(c-2)}$$

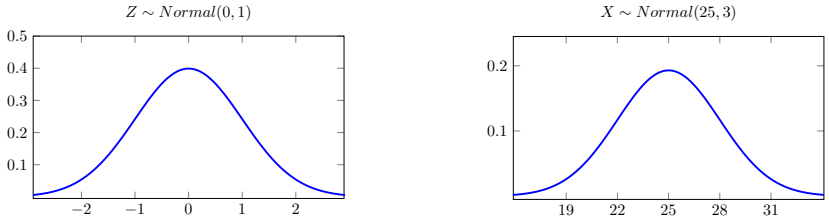
Cross-multiplying and solving the resulting quadratic, $2c(c-2) = 16$, the positive solution is $c = 4$, which we can substitute back into $4 = \frac{cm}{c-1}$ to obtain

$$4 = \frac{m \cdot 4}{3} \quad \to \quad m = 3.$$

4.7 The Normal Distribution

Definition 4.8 *The random variable X is normally distributed with parameters μ (the mean) and $\sigma > 0$ (the standard deviation) if its pdf is given by the function below on the interval $(-\infty, \infty)$.*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



This is also referred to as the Gaussian distribution, after mathematician Carl Friedrich Gauss. It will play a central role in the theory of statistics we'll develop in the next few chapters.

Remark Unfortunately, the pdf f_X for the normal distribution is not integrable in elementary terms. In other words, its antiderivative can't be expressed using only the functions and operations you've encountered in Cal II, which is a big problem, as it means we can't find areas under the curve using those methods.

The traditional way to get around this problem is to use a large table of values. Unless otherwise stated, the letter Z will be reserved for the random variable $Z \sim Normal(0,1)$ from now on, and a Z-table is simply a large table of values of the cdf $F_Z(z)$, that is, the area to the left of z, for various $z \in \mathbb{R}$.

What if X is a normally distributed random variable with a mean which is not zero and a standard deviation which is not one? As you might expect given the two graphs above, any two normal distributions differ by shifting and scaling, in the following precise sense.

Proposition 4.12 *$X \sim Normal(\mu, \sigma)$, then $Z = \frac{X-\mu}{\sigma} \sim Normal(0,1)$.*

Pf. Using substitution, let $z = \frac{x-\mu}{\sigma} = \frac{1}{\sigma}x - \frac{\mu}{\sigma}$, so $dz = \frac{1}{\sigma}\,dx$, and $dx = \sigma\,dz$.

$$\begin{aligned} f_X(x)\,dx &= \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\sigma\,dz \\ &= \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(\mu+\sigma z-\mu)^2}{2\sigma^2}}\,\sigma\,dz \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\,dz = f_Z(z)\,dz \end{aligned}$$

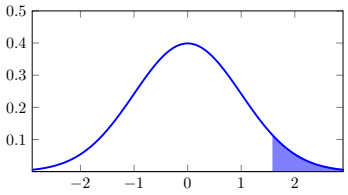
Thus, f_Z is the pdf of a normal distribution with $\mu = 0$ and $\sigma = 1$ as claimed. □

The process of translating values of any normally distributed random variable into corresponding values of Z is known as standardization. To standardize a value x of a normally distributed random variable $X \sim Normal(\mu, \sigma)$ we simply apply the formula $z = \frac{x-\mu}{\sigma}$, and we can treat the resulting z as a value of $Z \sim Normal(0, 1)$.

Example 4.18 *The height of a randomly selected Canadian adult female, in inches, is approximated by the random variable $X \sim Normal(64, 2.5)$. What proportion of Canadian adult females are taller than 5'8"? What proportion have heights between 5'3" and 5'7"?*

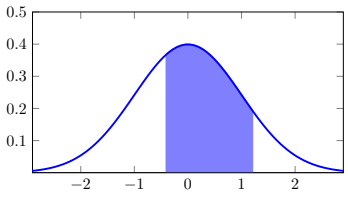
To standardize 5'8" = 68", we calculate $z = \frac{68-64}{2.5} = 1.6$. Then we have

$$\begin{aligned} P(X > 68) &= P(Z > 1.6) \\ &= 1 - P(Z \leq 1.6) \\ &= 1 - 0.9452 \\ &= 0.0548 \simeq 5.5\%. \end{aligned}$$



Thus, about 5.5% of Canadian adult females are taller than 5'8". For the second part we standardize 5'3" = 63" and 5'7" = 67" to obtain $z_1 = \frac{63-64}{2.5} = -0.4$ and $z_2 = \frac{67-64}{2.5} = 1.2$, then calculate

$$\begin{aligned} P(63 < X < 67) &= P(-0.4 < Z < 1.2) \\ &= P(Z \leq 1.2) - P(Z \leq -0.4) \\ &= 0.8849 - 0.3446 \\ &= 0.5404 \simeq 54\%. \end{aligned}$$

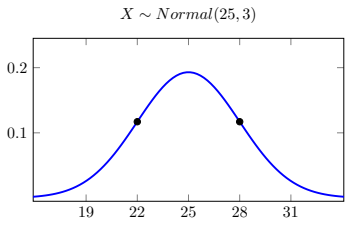


Therefore, around 54% of Canadian adult females are between 5'3" and 5'7".

Nowadays, we have easy access to computers which can calculate areas under the pdf of $Z \sim Normal(0, 1)$ efficiently using numerical integration, so the Z-table is something of a mathematical artifact whose uses are limited to classrooms and standardized tests.

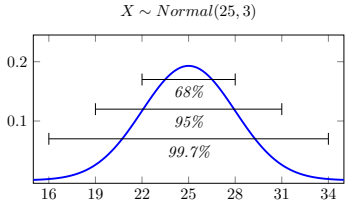
Normal Distribution Shortcuts

Any normal distribution has two inflection points, and these points occur exactly one standard deviation from the mean. This fact is useful when drawing a quick sketch of a normal distribution.



For approximate reasoning about normally distributed quantities without the aid of a Z-table or computer, the proposition below presents a well-known rule of thumb.

Proposition 4.13 *(68-95-99.7 Rule) For any normal distribution, 68% of the area under its pdf occurs within one standard deviation of the mean, 95% occurs within two standard deviations, and 99.7% occurs within three standard deviations.*



As well as being a useful rule of thumb for quick approximate calculations, the rule also illustrates how quickly the distribution decays when it moves away from the mean. For an arbitrary random variable, Chebyshev's Inequality implies that values less than two standard deviations from the mean must occur at least $1 - \frac{1}{2^2} = 75\%$ of the time, so values more than two standard deviations from the mean can occur up to 25% of the time. By the 68-95-99.7 Rule above, such values occur only 5% of the time in a normal distribution. In this sense, normally distributed quantities are very much concentrated around the mean.

Example 4.19 *If $X \sim Normal(64, 2.5)$ is used to model the height of a randomly selected Canadian female, provide a range of heights which will include 95% of all Canadian adult females.*

Using the 68-95-99.7 Rule, 95% of Canadian adult females will have heights within two standard deviations of the mean height, 64". So we calculate $64 - 2(2.5) = 59$ and $64 + 2(2.5) = 69$. Therefore, 95% of all Canadian adult females are taller than 59" = 4'11" and shorter than 69" = 5'9".

Chapter 5

Point Estimation

Introduction: The German Tank Problem

During the course of a war, three enemy tanks of a particular type are captured. Each tank has a serial number printed inside it. The first tank of this type to have been produced has serial number one, the second has serial number two, and so on.

The serial numbers found in the three tanks are 121, 53, and 246. What is your best guess at the total number of tanks of this type that have been produced?

The task of making a single guess at an unknown quantity, using information in a sample, is known as point estimation. Note that the point estimation problem above is very open-ended. What exactly does it mean to answer correctly?

5.1 Statistics & Parameters

Definition 5.1 *A parameter is a numerical property of a population being studied, and a statistic is a numerical property of a sample taken from this population.*

The key idea is to always distinguish is between values obtained from complete information, with no uncertainty, and those obtained from incomplete information via some sampling process.

Example 5.1 *If we wish to estimate the average age of all students in a classroom, the population in question is the set of all students in the room. Their average age is the parameter we want to estimate. If we ask a randomly selected group of five students their ages and average those, the value obtained is a statistic.*

Example 5.2 *In the German tank problem given in the introduction to this chapter, the population being studied is the set of all tanks of a particular type. Each*

serial number is represented once, so the distribution of serial numbers in the population is $Uniform(m)$, where m, the largest serial number, is the parameter we would like to estimate. Any value which can be derived from the given sample of three serial numbers we were given will be a statistic.

The central theme of inferential statistics is how to use statistics which we know to estimate parameters which we don't, and how to quantify the inherent error in our estimation methods. In many fields, sample information is the only information we have access to, and decisions must be made based on it. This makes inferential statistics a necessary tool over an extremely wide range of applications.

5.2 Sampling Distributions & Estimators

The value of a statistic is determined by the sample it was derived from, and hence *a statistic is a function whose input is a sample, and whose output is a number.*

If we fix a sample size, n, define a sample space Ω as the set of possible samples of that size, and assign probabilities to each possible sample, then any statistic that can be computed from a sample of size n is a random variable defined on Ω (recall Definition 3.1). This is the central idea of frequentist statistics: a parameter is simply a value which may or may not be known, but a *statistic is a random variable* and like any random variable, *it has a distribution.*

Definition 5.2 *The sampling distribution of a statistic, over samples of size n, is the distribution of possible values of that statistic.*

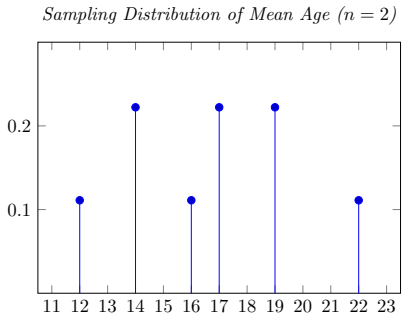
Example 5.3 *Consider a population of three individuals, Alice, Bob, and Charles. Alice is 12 years old, Bob is 16 years old, and Charles is 22 years old. Let $\overline{X}$ denote the mean age in a sample of two individuals drawn from this population ($\overline{X}$ is standard notation for a sample mean in statistical practice). Find the sampling distribution of $\overline{X}$ over all samples of size two taken with replacement.*

If we let X_1 and X_2 denote the ages of the first and second individuals selected in our sample, then $\overline{X} = \frac{1}{2}(X_1 + X_2)$. There are $3 \cdot 3 = 9$ possible samples, which we enumerate below and for each, calculate the value of $\overline{X}$.

Sample	X_1	X_2	$\overline{X}$
A,A	12	12	12
A,B	12	16	14
A,C	12	22	17
B,A	16	12	14
B,B	16	16	16
B,C	16	22	19
C,A	22	12	17
C,B	22	16	19
C,C	22	22	22

If we assume every sample is equally likely, then each sample above occurs with probability $\frac{1}{9}$, so we obtain the sampling distribution below.

$\overline{X}$	$P(\overline{X} = \overline{x})$
12	$\frac{1}{9}$
14	$\frac{2}{9}$
16	$\frac{1}{9}$
17	$\frac{2}{9}$
19	$\frac{2}{9}$
22	$\frac{1}{9}$



Random Sampling

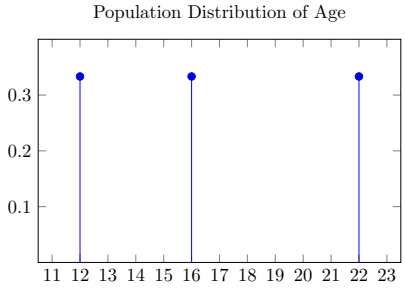
To compute the sampling distribution above, we assumed every sample was equally likely. The sampling distribution depends on the sampling process, so we must be explicit about what constitutes a sample and how samples are selected.

Definition 5.3 *(SRS Informally) If an ordered sequence of individuals is selected in a manner where every member of the population is equally likely to be chosen for each position in the sequence, we obtain a simple random sample (SRS) taken with replacement.*

Remark SRSs can also be taken without replacement. Sampling with replacement and without replacement are both common in practice, and the most important results we'll derive apply when sampling without replacement as well.

In Example 5.3 above, consider the distribution of age in the population. There is a single 12 year old, a single 16 year old, and a single 22 year old, so if X denotes the age of an individual selected uniformly at random from the population, then the distribution of X is given below.

$$f_X(x) = \begin{cases} \frac{1}{3} & \text{if } x = 12 \\ \frac{1}{3} & \text{if } x = 16 \\ \frac{1}{3} & \text{if } x = 22 \\ 0 & \text{otherwise} \end{cases}$$



In a SRS (with replacement) of two individuals taken from this population, the first individual chosen must be equally likely to be Alice, Bob, or Charles, the second

individual chosen must also be equally likely to be Alice, Bob, or Charles, and so on. If X_i represents the age of the i^{th} individual in our sample, then X_i must have the distribution given above, and moreover, knowledge of the value of X_i should not influence our beliefs about any X_j for $i \neq j$.

Definition 5.4 *(SRS Formally) Given a population distribution X, a simple random sample taken without replacement from X is a sequence $X_1, X_2, \ldots, X_n$ of independent, identically distributed random variables, each of which has the same distribution as X.*

Example 5.4 *Consider a population of four individuals, one has shoe size 7, one has shoe size 10, and the other two have shoe size 11. If a SRS of two individuals is taken with replacement from this population, what is the probability the second individual chosen has shoe size 11? If the first individual chosen has shoe size 10, what is the probability the second individual chosen has shoe size 11?*

Let X denote the distribution of shoe sizes in the population, so $P(X = 7) = \frac{1}{4}$, $P(X = 10) = \frac{1}{4}$, and $P(X = 11) = \frac{1}{2}$.

Our sample X_1, X_2 is a SRS taken with replacement, so the probability any X_i is a 11 should be $\frac{1}{2}$, regardless of the values of any other X_j. Thus, $P(X_2 = 11) = P(X = 11) = \frac{1}{2}$ and $P(X_2 = 11 \mid X_1 = 10) = P(X_2 = 11) = \frac{1}{2}$.

Estimators

When a statistic is being used to estimate a parameter, we call it an estimator for that parameter. Parameters are typically denoted with greek letters, such as θ, and the notation $\widehat{\theta}$ is often used to denote a statistic estimating θ.

Example 5.5 *Suppose that in the context of Example 5.3, the mean age in a SRS of size two is used to estimate the mean age in the population. The SRS selected consists of Alice and Charles. What are μ and $\widehat{\mu}$?*

The greek letter μ is typically used to denote a population mean, so in this scenario, we have the parameter $\mu = \frac{50}{3} \simeq 16.67$ and the estimator $\widehat{\mu} = \frac{1}{2}(X_1 + X_2)$ which with our sample, gives the estimate $\widehat{\mu} = \frac{1}{2}(12 + 22) = 17$.

In this case, the value of our estimator ended up very close to μ, which is exactly what we're hoping for. In practice though, we don't know the value of the parameter being estimated, or we wouldn't have been estimating it in the first place! Without this knowledge, how can we begin to evaluate the usefulness of an estimator?

Notice that the sampling distribution of the sample mean $\overline{X}$ that we constructed in Example 5.3 has a very nice property. If we compute its expected value

$$E(\overline{X}) = 12 \cdot \tfrac{1}{9} + 14 \cdot \tfrac{2}{9} + 16 \cdot \tfrac{1}{9} + 17 \cdot \tfrac{2}{9} + 19 \cdot \tfrac{2}{9} + 22 \cdot \tfrac{1}{9} = \tfrac{150}{9} = \tfrac{50}{3}$$

the result is precisely equal to the population mean μ. Thus, if we use the estimator $\widehat{\mu} = \overline{X} = \frac{1}{n}(X_1 + X_2)$ to estimate the population mean, then the average value of our estimator is equal to the parameter μ being estimated. We say that $\widehat{\mu} = \overline{X}$ is an unbiased estimator of μ.

Definition 5.5 *The bias of an estimator is the difference between its expected value and the actual value of the parameter being estimated.*

$$\boxed{Bias(\widehat{\theta}) = E(\widehat{\theta}) - \theta}$$

If $Bias(\widehat{\theta}) > 0$, the mean value of $\widehat{\theta}$ is an overestimate of θ, if $Bias(\widehat{\theta}) < 0$, the mean value of $\widehat{\theta}$ is an underestimate of θ, and if $Bias(\widehat{\theta}) = 0$ then $E(\widehat{\theta}) = \theta$ and $\widehat{\theta}$ is an unbiased estimator of θ.

The concept of bias is fairly intuitive. Suppose you estimate the average age of all students in a classroom by finding the oldest student in the class and reporting their age. This estimator will have positive bias. Similarly, if you estimate the average age of all students in a classroom by finding the youngest student in the class and reporting their age, that estimator will have negative bias.

Theorem 5.1 *In any population, the mean of a SRS taken with replacement is an unbiased estimator of the population mean μ.*

Pf. Let $\widehat{\mu} = \overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$, where X_i is the i^{th} value in our sample.

If our sample is a SRS taken with replacement, the distribution of each X_i is the same as the population distribution. Therefore, $E(X_i) = E(X) = \mu$, the population mean, and by linearity of expectation,

$$\begin{aligned} E(\widehat{\mu}) &= E(\tfrac{1}{n}(X_1 + X_2 + ... + X_n)) \\ &= \tfrac{1}{n}E(X_1 + X_2 + ... + X_n) \\ &= \tfrac{1}{n}(E(X_1) + E(X_2) + ... + E(X_n)) \\ &= \tfrac{1}{n}(\mu + \mu + ... + \mu) \\ &= \tfrac{1}{n}(n\mu) \\ &= \mu. \end{aligned}$$

□

Example 5.6 *Consider a miniature version of the German Tank Problem given in the introduction to this chapter, where the population consists of only four tanks, with serial numbers 1 through 4. Let θ denote the population maximum, so $\theta = 4$.*

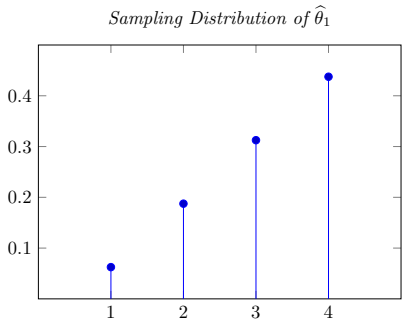
If we estimate θ by taking the maximum serial number in a sample of size two, show we'll obtain a biased estimate. If we estimate θ by doubling the average serial number in a sample of size two then subtracting one, show we'll obtain an unbiased estimate.

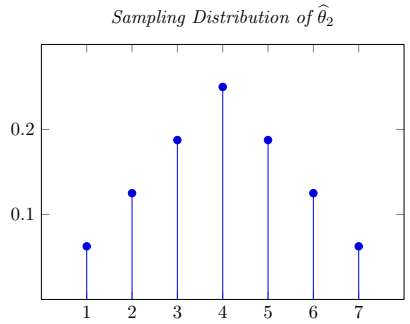
Define $\widehat{\theta}_1 = max(X_1, X_2)$ *and* $\widehat{\theta}_2 = 2 \cdot \frac{1}{2}(X_1 + X_2) - 1$. *There are sixteen possible samples taken with replacement, so we have the samples and corresponding values of the two estimators given below.*

X_1	X_2	$\widehat{\theta}_1$	$\widehat{\theta}_2$
1	1	1	1
1	2	2	2
1	3	3	3
1	4	4	4
2	1	2	2
2	2	2	3
2	3	3	4
2	4	4	5
3	1	3	3
3	2	3	4
3	3	3	5
3	4	4	6
4	1	4	4
4	2	4	5
4	3	4	6
4	4	4	7

x	$P(\widehat{\theta}_1 = x)$
1	$\frac{1}{16}$
2	$\frac{3}{16}$
3	$\frac{5}{16}$
4	$\frac{7}{16}$

x	$P(\widehat{\theta}_2 = x)$
1	$\frac{1}{16}$
2	$\frac{2}{16}$
3	$\frac{3}{16}$
4	$\frac{4}{16}$
5	$\frac{3}{16}$
6	$\frac{2}{16}$
7	$\frac{1}{16}$





Now we can compute the expected values and evaluate the bias of each estimator.

$$E(\widehat{\theta}_1) = 1 \cdot \tfrac{1}{16} + 2 \cdot \tfrac{3}{16} + 3 \cdot \tfrac{5}{16} + 4 \cdot \tfrac{7}{16} = \tfrac{50}{16} = \tfrac{25}{8}$$
$$E(\widehat{\theta}_2) = 1 \cdot \tfrac{1}{16} + 2 \cdot \tfrac{2}{16} + 3 \cdot \tfrac{3}{16} + 4 \cdot \tfrac{4}{16} + 5 \cdot \tfrac{3}{16} + 6 \cdot \tfrac{2}{16} + 7 \cdot \tfrac{1}{16} = \tfrac{64}{16} = 4$$

Thus, $Bias(\widehat{\theta}_1) = \frac{25}{8} - 4 = -\frac{7}{8} = -0.875$ *and* $Bias(\widehat{\theta}_2) = 4 - 4 = 0$.

Remark If the population we're studying is small, we can explicitly compute the sampling distribution by writing down every possible sample, as above. Of course, this is rarely feasible in practice.

In general, the sampling distribution can be approximated by repeatedly taking random samples using a computer and recording the values of the relevant statistic. The distribution obtained will be a better and better approximation of the true sampling distribution as the number of samples grows larger and larger.

5.3 Evaluating Estimators

What determines whether an estimator produces useful estimates of a parameter, and what makes one estimator better than another? Clearly, good estimators should take values close to the parameter being estimated, so the smaller that the typical difference between the two is, the better.

Definition 5.6 *The mean-squared error of an estimator is the expected value of the squared difference between the estimator and the parameter it estimates.*

$$\boxed{MSE(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2]}$$

Example 5.7 *Find the MSE of the two estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ in Example 5.6.*

In both cases, the parameter being estimated is $\theta = 4$, and from the sampling distributions of $\widehat{\theta}_1$ and $\widehat{\theta}_2$, we can compute

$$\begin{aligned}
MSE(\widehat{\theta}_1) &= E[(\widehat{\theta}_1 - \theta)^2] \\
&= (1-4)^2 \cdot \tfrac{1}{16} + (2-4)^2 \cdot \tfrac{3}{16} + (3-4)^2 \cdot \tfrac{5}{16} + (4-4)^2 \cdot \tfrac{7}{16} \\
&= 9 \cdot \tfrac{1}{16} + 4 \cdot \tfrac{3}{16} + 1 \cdot \tfrac{5}{16} + 0 \cdot \tfrac{7}{16} = \tfrac{26}{16} \\
MSE(\widehat{\theta}_2) &= E[(\widehat{\theta}_2 - \theta)^2] \\
&= (1-4)^2 \cdot \tfrac{1}{16} + (2-4)^2 \cdot \tfrac{2}{16} + (3-4)^2 \cdot \tfrac{3}{16} + \ \ldots \ + (7-4)^2 \cdot \tfrac{1}{16} \\
&= 9 \cdot \tfrac{1}{16} + 4 \cdot \tfrac{2}{16} + 1 \cdot \tfrac{3}{16} + 0 \cdot \tfrac{4}{16} + 1 \cdot \tfrac{3}{16} + 4 \cdot \tfrac{2}{16} + 9 \cdot \tfrac{1}{16} = \tfrac{40}{16}
\end{aligned}$$

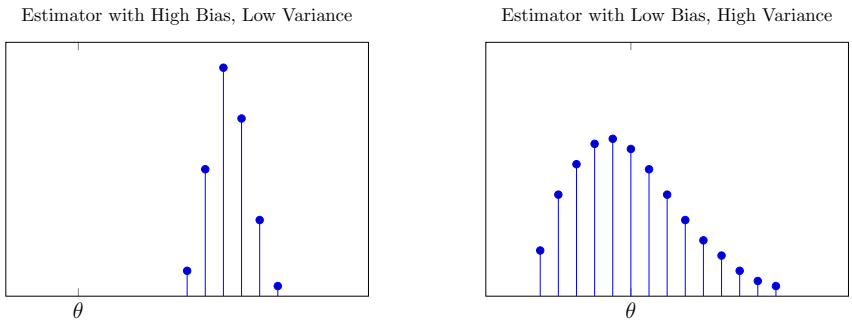
Theorem 5.2 *$MSE(\widehat{\theta}) = Bias^2(\widehat{\theta}) + Var(\widehat{\theta})$, where $Var(\widehat{\theta}) = E[(\widehat{\theta} - E(\widehat{\theta}))^2]$*

Pf. The result follows from a calculation using linearity of expectation.

$$\begin{aligned}
\text{MSE}(\widehat{\theta}) &= E[(\widehat{\theta} - \theta)^2] \\
&= E[(\widehat{\theta} - E[\widehat{\theta}] + E[\widehat{\theta}] - \theta)^2] \\
&= E[(\widehat{\theta} - E[\widehat{\theta}])^2 + 2(\widehat{\theta} - E[\widehat{\theta}])(E[\widehat{\theta}] - \theta) + (E[\widehat{\theta}] - \theta)^2] \\
&= E[(\widehat{\theta} - E[\widehat{\theta}])^2] + E[2(\widehat{\theta} - E[\widehat{\theta}])(E[\widehat{\theta}] - \theta)] + E[(E[\widehat{\theta}] - \theta)^2] \\
&= E[(\widehat{\theta} - E[\widehat{\theta}])^2] + E[2(\widehat{\theta} - E[\widehat{\theta}])(E[\widehat{\theta}] - \theta)] + (E[\widehat{\theta}] - \theta)^2 \\
&= \text{Var}(\widehat{\theta}) + E[2(\widehat{\theta} - E[\widehat{\theta}])\text{Bias}(\widehat{\theta})] + \text{Bias}^2(\widehat{\theta}) \\
&= \text{Var}(\widehat{\theta}) + 2\text{Bias}(\widehat{\theta}) \cdot E[\widehat{\theta} - E(\widehat{\theta})] + \text{Bias}^2(\widehat{\theta}) \\
&= \text{Var}(\widehat{\theta}) + 2\text{Bias}(\widehat{\theta}) \cdot (E[\widehat{\theta}] - E[E(\widehat{\theta})]) + \text{Bias}^2(\widehat{\theta}) \\
&= \text{Var}(\widehat{\theta}) + 2\text{Bias}(\widehat{\theta}) \cdot (E[\widehat{\theta}] - E[\widehat{\theta}]) + \text{Bias}^2(\widehat{\theta}) \\
&= \text{Bias}^2(\widehat{\theta}) + \text{Var}(\widehat{\theta})
\end{aligned}$$

□

The bias of an estimator measures the distance between the mean of the sampling distribution and the parameter being estimated, while the variance measures the dispersion in the sampling distribution.



Consider as an analogy an archery contest in which five shots are taken by each competitor. Estimators with high bias and low variance are akin to archers whose five arrows fall very close together, but all land quite far from the target. Estimators with low bias and high variance are akin to archers whose arrows are centred around the middle of the target, but with each individual arrow potentially landing quite far from it.

The best estimators are those whose values centre around the parameter being estimated and also vary as little as possible. That is, those with low bias and low variance. Theorem 5.2 assures us that an estimator with both these properties will also have a small mean-squared error.

Example 5.8 *Consider the Mini German Tank Problem in Example 5.6. Find the mean-squared error, bias, and variance of the estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ whose sampling distributions we calculated.*

We already have $MSE(\widehat{\theta}_1) = \frac{26}{26}$ and $Bias(\widehat{\theta}_1) = -\frac{7}{8}$ from our earlier calculations, so by Theorem 5.2, we have

$$\begin{aligned} MSE(\widehat{\theta}_1) &= Bias^2(\widehat{\theta}_1) + Var(\widehat{\theta}_1) \\ \tfrac{26}{16} &= (-\tfrac{7}{8})^2 + Var(\widehat{\theta}_1) \\ Var(\widehat{\theta}_1) &= \tfrac{55}{64}. \end{aligned}$$

Since $\widehat{\theta}_2$ is unbiased, $MSE(\widehat{\theta}_2) = Var(\widehat{\theta}_2)$ by Theorem 5.2, so $Var(\widehat{\theta}_2) = \frac{40}{16}$. Note that $\widehat{\theta}_1$ is the estimator with the lower MSE despite the fact that it's biased.

5.4 Maximum Likelihood Estimation

One strategy for estimating a parameter from sample data is to determine the value that makes the observed sample as likely as possible to occur. If I told you that in a sample of 50 flips of a biased coin, 40 heads and 10 tails occurred, then $\frac{40}{50} = 0.8$ clearly seems like the best guess you could make for the probability of heads on a single flip of that coin. This value is called a maximum likelihood estimate.

Definition 5.7 *Suppose we have a SRS taken with replacement $x_1, x_2, \dots, x_n$ from a population distribution X with an unknown parameter θ. The likelihood function for this sample is denoted $\mathcal{L}(x_1, x_2, \dots, x_n \,|\, \theta)$ and defined by*

$$\boxed{\mathcal{L}(x_1, x_2, \dots, x_n \,|\, \theta) = f_X(x_1) f_X(x_2) \dots f_X(x_n) = \prod_{i=1}^{n} f_X(x_i)}$$

Since our sample is a SRS, $f_X(x_1) f_X(x_2) \dots f_X(x_n)$ is the probability (or probability density) of observing the sample values $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$. The maximum likelihood estimate is the value of θ that maximizes $\mathcal{L}(x_1, x_2, \dots, x_n \,|\, \theta)$.

We can use familiar techniques from differential calculus to find the maximum. One useful trick to recall is that because $\ln(x)$ is an increasing function, the value of θ that maximizes $\mathcal{L}(x_1, x_2, \dots, x_n \,|\, \theta)$ will also maximize $\ln(\mathcal{L}(x_1, x_2, \dots, x_n \,|\, \theta))$. This will greatly simplify calculations since the logarithm of a product can be written as a sum of logarithms, and derivatives of sums are much easier to deal with. To take advantage of this, let's denote the log-likelihood with $\ell(x_1, x_2, \dots, x_n \,|\, \theta)$.

$$\boxed{\ell(x_1, x_2, \dots, x_n \,|\, \theta) = \ln\left(f_X(x_1) f_X(x_2) \dots f_X(x_n)\right) = \sum_{i=1}^{n} \ln\left(f_X(x_i)\right)}$$

Example 5.9 *Let $X \sim$ Exponential(λ), and suppose we observe a SRS of three values from this distribution with $X_1 = 1, X_2 = 3, X_3 = 4$. Find the maximum likelihood estimate of λ.*

Note that $\lambda \in (0, \infty)$ for an exponential distribution. To begin with, we'll write out the log-likelihood and simplify.

$$\begin{aligned}
\ell(x_1, x_2, x_3 \,|\, \lambda) &= \ln\left(f_X(x_1) f_X(x_2) f_X(x_3)\right) \\
&= \ln\left(f_X(x_1)\right) + \ln\left(f_X(x_2)\right) + \ln\left(f_X(x_3)\right) \\
&= \ln\left(f_X(1)\right) + \ln\left(f_X(3)\right) + \ln\left(f_X(4)\right) \\
&= \ln(\lambda e^{-1\lambda}) + \ln(\lambda e^{-3\lambda}) + \ln(\lambda e^{-4\lambda}) \\
&= \ln(\lambda) + \ln(e^{-1\lambda}) + \ln(\lambda) + \ln(e^{-3\lambda}) + \ln(\lambda) + \ln(e^{-4\lambda}) \\
&= 3\ln(\lambda) - (\lambda + 3\lambda + 4\lambda) \\
&= 3\ln(\lambda) - 8\lambda
\end{aligned}$$

The point of all this simplifying is to make computing the derivative $\frac{d\ell}{d\lambda}$ as easy as possible. Once we have the derivative, we'll solve $\frac{d\ell}{d\lambda} = 0$ to find the critical points.

$$\frac{d\ell}{d\lambda} = \frac{3}{\lambda} - 8 \quad \to \quad 0 = \frac{3}{\lambda} - 8 \quad \to \quad \lambda = \frac{3}{8}$$

Thus, $\lambda = \frac{3}{8}$ is the only critical point on $(0, \infty)$. Furthermore, $\frac{d^2\ell}{d\lambda^2} = -\frac{3}{\lambda^2}$, which is always negative, hence by the second derivative test, a maximum occurs at $\lambda = \frac{3}{8}$. Therefore, the maximum likelihood estimate for the parameter λ is $\widehat{\lambda} = \frac{3}{8}$.

Note that we can easily generalize this result for any sample values. The key point is that the observed sample values $x_1, x_2, .., x_n$ can be treated as constants in the derivative calculation.

Example 5.10 *Let $X \sim Exponential(\lambda)$. Suppose we observe a SRS of n values $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$. Find the maximum likelihood estimator for λ.*

As above, note $\lambda \in (0, \infty)$, and we can simplify the log-likelihood as below.

$$\begin{aligned}\ell(x_1, x_2, ..., x_n \,|\, \lambda) &= \ln\left(f_X(x_1) f_X(x_2) ... f_X(x_n)\right) \\ &= \ln\left(f_X(x_1)\right) + \ln\left(f_X(x_2)\right) + \ ... \ + \ln\left(f_X(x_n)\right) \\ &= \ln(\lambda e^{-\lambda x_1}) + \ln(\lambda e^{-\lambda x_2}) + \ ... \ + \ln(\lambda e^{-\lambda x_n}) \\ &= \ln(\lambda) + \ln(e^{-\lambda x_1}) + \ln(\lambda) + \ln(e^{-\lambda x_2}) + \ln(\lambda) + \ ... \ \ln(e^{-\lambda x_n}) \\ &= n \ln(\lambda) - (\lambda x_1 + \lambda x_2 + \ ... \ + \lambda x_n)\end{aligned}$$

Now we compute $\frac{d\ell}{d\lambda}$ and find the critical points.

$$\frac{d\ell}{d\lambda} = \frac{n}{\lambda} - (x_1 + x_2 + \ ... \ + x_n) \quad \rightarrow \quad 0 = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \quad \rightarrow \quad \lambda = \frac{n}{\sum_{i=1}^{n} x_i}$$

The same argument as above shows this critical point yields the global maximum, so the maximum likelihood estimator for λ is $\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i}$. Note that this is simply the reciprocal of the sample mean.

Example 5.11 *If $X \sim Uniform(m)$, find the maximum likelihood estimator of m from a SRS $X_1 = x_1, X_2 = x_2, \ ..., X_n = x_n$ taken with replacement.*

In this case the likelihood function is easy to deal with because the pmf for a uniform distribution is very simple. Any possible sample of n values has the same likelihood.

$$\begin{aligned}\mathcal{L}(x_1, x_2, ..., x_n \,|\, m) &= f_X(x_1) f_X(x_2) ... f_X(x_n) \\ &= \frac{1}{m} \cdot \frac{1}{m} \cdot \ ... \ \cdot \frac{1}{m} = \frac{1}{m^n}\end{aligned}$$

Note that we are assuming each $x_i \leq m$ in the calculation above. If $x_i > m$ then $f_X(x_i) = 0$ since such an x_i will never appear in any sample. Thus, the likelihood function has the form below.

$$\mathcal{L}(x_1, x_2, ..., x_n \,|\, m) = \begin{cases} \frac{1}{m^n} & \text{if } \ m \geq x_i \text{ for } 1 \leq i \leq n \\ 0 & \text{otherwise} \end{cases}$$

Since the sample size n is fixed, to maximize this function we need m to be as small as possible, while obeying the constraint that $m \geq x_i$ for $1 \leq i \leq n$. In other words, the maximum likelihood estimator is the smallest value for m which is larger than or equal to each of the observed sample values. Thus, $\widehat{m} = max(X_1, X_2, ..., X_n)$.

We've seen a concrete example involving this estimator already, Example 5.6. In that context, $\widehat{\theta_1}$ was the maximum likelihood estimator of $\theta = 4$. The bias of $\widehat{\theta_1}$ was nonzero, and this demonstrates that the maximum likelihood estimator is not, in general, unbiased.

Properties of the Maximum Likelihood Estimator

Aside from the logic that lead us to consider the maximum likelihood estimator in the first place (our best guess for θ should be the value that makes the data we've actually observed as likely as possible), the maximum likelihood estimator also has some nice properties that give it additional appeal.

Although the maximum likelihood estimator is not unbiased, it is asymptotically unbiased, meaning that as the sample size grows very large, its bias converges to zero. In addition, it also has asymptotically minimal variance, which means that as the sample size grows very large, its variance will eventually be smaller than or equal to the variance of any other unbiased estimator.

Proposition 5.1 *As the sample size $n \to \infty$, $Bias(\widehat{\theta}_{MLE}) \to 0$ and if $\widehat{\theta}_U$ is any other unbiased estimator of θ, then $Var(\widehat{\theta}_{MLE}) \leq Var(\widehat{\theta}_U) + \epsilon$ where $\epsilon \to 0$.*

This proposition gives us confidence that the maximum likelihood estimator will do a better and better job of estimating a parameter as the amount of sample data we acquire grows larger and larger. The proof is beyond the scope of this course, but the idea of asymptotic unbiasedness is illustrated in the example below.

Example 5.12 *Let $X \sim Uniform(m)$. In Example 5.11, we saw the maximum likelihood estimator for the parameter m is $\widehat{m} = max(X_1, X_2, ..., X_n)$. Show that $\widehat{m}$ is asymptotically unbiased.*

Consider the expected value of the estimator $\widehat{m}$,

$$E(\widehat{m}) = 1 \cdot P(\widehat{m} = 1) + 2 \cdot P(\widehat{m} = 2) + \ ... \ + m \cdot P(\widehat{m} = m).$$

Observe that $\widehat{m} = m$ iff the value m appears in our sample. In an SRS taken with replacement from $X \sim Uniform(m)$, each sample value X_i has $P(X_i = m) = \frac{1}{m}$.

$$\begin{aligned} P(\widehat{m} = m) &= P((X_1 = m) \cup (X_2 = m) \cup ... \cup (X_n = m)) \\ &= 1 - P((X_1 \neq m) \cap (X_2 \neq m) \cap ... \cap (X_n \neq m)) \\ &= 1 - (P(X_1 \neq m) \cdot P(X_2 \neq m) \cdot \ ... \ \cdot P(X_n \neq m)) \\ &= 1 - \tfrac{m-1}{m} \cdot \tfrac{m-1}{m} \cdot \ ... \ \cdot \tfrac{m-1}{m} \\ &= 1 - (\tfrac{m-1}{m})^n \end{aligned}$$

Thus, $P(\widehat{m} = m) \to 1$ as $n \to \infty$, because $\frac{m-1}{m} < 1$. Therefore $E(\widehat{m}) \to m \cdot 1 = m$. We conclude that $\widehat{m}$ is asymptotically unbiased since $Bias(\widehat{m}) = E(\widehat{m}) - m \to m - m = 0$ as $n \to \infty$.

Another useful fact concerning maximum likelihood estimation is that if one parameter is a function of another, then applying that function to the maximum likelihood estimator of the former parameter will yield the maximum likelihood estimator of the latter.

Proposition 5.2 *Suppose θ and ψ are two parameters of a distribution, and denote the maximum likelihood estimators of these by parameters by $\widehat{\theta}_{MLE}$ and $\widehat{\psi}_{MLE}$. If $\psi = g(\theta)$, then $\widehat{\psi}_{MLE} = g(\widehat{\theta}_{MLE})$.*

Example 5.13 *Let $X \sim Uniform(m)$, and let μ denote the mean of X. Find the maximum likelihood estimator for μ.*

In Example 5.11, we found that $\widehat{m}_{MLE} = max(X_1, X_2, \dots, X_n)$. Since $\mu = \frac{m+1}{2}$ in a discrete uniform distribution, we have $\widehat{\mu}_{MLE} = \frac{\widehat{m}_{MLE}+1}{2} = \frac{max(X_1,X_2,\dots,X_n)+1}{2}$.

Example 5.14 *Find the maximum likelihood estimator for the standard deviation σ of $X \sim Bernoulli(p)$.*

For a Bernoulli distribution, $\sigma = \sqrt{p(1-p)}$ by Theorem 3.7, so if we can find the maximum likelihood estimator for p, then $\widehat{\sigma}_{MLE} = \sqrt{\widehat{p}_{MLE}(1-\widehat{p}_{MLE})}$.

Note that in a SRS with values $x_1, x_2, \dots, x_n$ drawn from a Bernoulli distribution, every value is either zero (with probability $1-p$) or one (with probability p). Let k denote the number of ones in our sample, then

$$\begin{aligned}
&\ell(x_1, x_2, \dots, x_n \,|\, p) = \ln(f_X(x_1)) + \dots + \ln(f_X(x_n)) \\
&= \ln(p^{x_1}(1-p)^{1-x_1}) + \dots + \ln(p^{x_n}(1-p)^{1-x_n})) \\
&= x_1 \ln(p) + (1-x_1)\ln(1-p) + \dots + x_n \ln(p) + (1-x_n)\ln(1-p) \\
&= \ln(p)\sum_{i=1}^{n} x_i + \ln(1-p)\sum_{i=1}^{n}(1-x_i)
\end{aligned}$$

The log-likelihood is well-defined and differentiable for $p \in (0,1)$, so we can maximize it by computing $\frac{d\ell}{dp}$ and identifying the critical points. Since $\lim_{x\to 0^+} \ln(x) = -\infty$ and $\ln(1) = 0$, the log-likelihood tends to $-\infty$ as p approaches either endpoint of $(0,1)$. Thus, if there's a single critical point on $(0,1)$, it must be a max.

$$\begin{aligned}
\frac{d\ell}{dp} &= \frac{1}{p}\sum_{i=1}^{n} x_i + \frac{1}{1-p}(-1)\sum_{i=1}^{n}(1-x_i) \\
0 &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\sum_{i=1}^{n} 1 - \sum_{i=1}^{n} x_i}{1-p} \\
\frac{n - \sum_{i=1}^{n} x_i}{1-p} &= \frac{\sum_{i=1}^{n} x_i}{p} \\
\frac{n-S}{1-p} &= \frac{S}{p} \quad \textit{where} \quad S = \sum_{i=1}^{n} x_i
\end{aligned}$$

Cross multiplying and solving for p, we obtain $np - Sp = S - Sp$ *and* $p = \frac{S}{n}$*, which is simply the proportion of ones in our sample.*

We conclude that $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *is the maximum likelihood estimator for p, so by Proposition 5.2, the maximum likelihood estimator for* σ *is*

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i \left(1 - \frac{1}{n}\sum_{i=1}^{n} X_i\right)}.$$

Remark The log-likelihood above is undefined when $p = 0$ or $p = 1$, but these are possible values for the parameter p if $X \sim Bernoulli(p)$. Despite this, note that the maximum likelihood estimator for p will estimate $\widehat{p} = 0$ if no successes are observed in the sample, and will estimate $\widehat{p} = 1$ if every sample value is a success.

Chapter 6

Interval Estimation

Introduction: Quantifying Estimation Error

The techniques of point estimation which we've developed over the last chapter are useful when a single well justified guess at the value of an unknown parameter is called for. However, point estimates often do not provide sufficient information to inform important decisions.

Suppose a report on labor in Canada states that the unemployment rate has risen from 6.8% in June 2016 to 7.1% in June 2017. Each of these numbers is a point estimate derived from a sample, and therefore each represents a single observation drawn from a sampling distribution.

The naive interpretation of the report on labor is that the unemployment rate has gone up, but those two figures stated depend not only on the actual state of the labor market, but also on a random result of a sampling process. It's possible that the unemployment rate in Canada in June 2016 was in fact 7.0%, we just happened to draw a sample which resulted in an underestimate, and perhaps the unemployment rate in Canada in June 2017 was in fact 6.9%, we just happened to draw a sample which resulted in an overestimate.

When providing an estimate of an unknown parameter, we would like to be able to quantify the uncertainty inherent in the process that produced it. In our example, instead of stating (probably falsely) that the unemployment rate in Canada in June 2016 was 6.8%, it would be much more useful to say that with 95% confidence, the unemployment rate in Canada in June 2016 was between, say, 6.6% and 7.0%. This way when we compare two statistics, we'll be better able to assess whether the difference between those two values is significant or not.

6.1 The Law of Large Numbers

The law of large numbers is one of the cornerstones of statistical inference. This is a mathematically precise statement of the fairly intuitive principle that if we take a larger and larger SRS of values, the mean in our sample tends to become closer and closer to the mean in the population. This principle is often informally referred to as 'regression to the mean'.

Theorem 6.1 *(Weak Law of Large Numbers) Suppose that X is a random variable with mean μ and standard deviation σ. If $X_1, X_2, \ldots, X_n$ is a SRS of values of X taken with replacement, then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\right| \geq \epsilon\right) = 0$$

In other words, for any interval containing μ, no matter how tiny it might be, the probability of obtaining a sample whose mean is outside that interval approaches zero as the sample size approaches infinity. This result holds for any population distribution whose expected value and standard deviation are defined.

Pf. Let $\overline{X} = \frac{X_1+X_2+\ldots+X_n}{n}$. Our sample $X_1, X_2, \ldots, X_n$ is a SRS, so each X_i has the same distribution as X, hence $E(X_i) = E(X) = \mu$. By linearity of expectation,

$$\begin{aligned} E(\overline{X}) &= E\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) \\ &= \frac{1}{n}(E(X_1) + E(X_2) + \ldots + E(X_n)) \\ &= \frac{1}{n}(n\mu) = \mu. \end{aligned}$$

We can also obtain a simple expression for the variance of $\overline{X}$ by making use of the properties of variance we demonstrated in Theorem 3.6 and 3.5.

$$\begin{aligned} \mathrm{Var}(\overline{X}) &= \mathrm{Var}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) \\ &= \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \ldots + X_n) \\ &= \frac{1}{n^2}(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n)) \\ &= \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n} \end{aligned}$$

Thus, the mean of $\overline{X}$ is μ, and the standard deviation of $\overline{X}$ is $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$.

Our goal is to show that $\overline{X}$ can't stray too far from it's mean, μ, and we have a tool for bounding the values a random variable can take, Chebyshev's Inequality (Theorem 4.2). Applying the inequality to $\overline{X}$,

$$P\big(\overline{X} \in (\mu - k\tfrac{\sigma}{\sqrt{n}},\, \mu + k\tfrac{\sigma}{\sqrt{n}})\big) \geq 1 - \frac{1}{k^2}$$
$$P\big(\overline{X} \notin (\mu - k\tfrac{\sigma}{\sqrt{n}},\, \mu + k\tfrac{\sigma}{\sqrt{n}})\big) \leq \frac{1}{k^2}.$$

Let $\epsilon = k\frac{\sigma}{\sqrt{n}}$, so we have $k = \frac{\epsilon\sqrt{n}}{\sigma}$ and $\frac{1}{k^2} = \frac{\sigma^2}{\epsilon^2 n}$. We can rewrite the inequality above as

$$P\big(\overline{X} \notin (\mu - \epsilon,\, \mu + \epsilon)\big) \leq \frac{\sigma^2}{\epsilon^2 n}$$
$$P\left(\left|\frac{X_1 + X_2 + \ ... \ + X_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2 n}.$$

To complete the argument, observe that $\frac{\sigma^2}{\epsilon^2 n}$ approaches zero as $n \to \infty$. □

The results concerning the mean and standard deviation of $\overline{X}$ we obtained before applying Chebyshev's Inequality are significant in their own right, so we'll restate them below.

Corollary 6.1 *Suppose $X_1, X_2, \ldots, X_n$ is a SRS drawn from a population with mean μ and standard deviation σ, and let $\overline{X} = \frac{X_1, X_2, \ldots, X_n}{n}$. The mean and standard deviation of the sampling distribution of $\overline{X}$ are given by*

$$\boxed{\mu_{\overline{X}} = \mu \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}}$$

The strong law of large numbers, given below, is also a statement that the mean of a SRS will converge to the population mean as the sample size approaches infinity, but the notion of convergence is slightly different. This alternate formulation is more difficult to prove, but can be easier to apply.

Theorem 6.2 *(Strong Law of Large Numbers) Suppose that X is a random variable with mean μ. If $X_1, X_2, \ldots, X_n$ is a SRS of values of X taken with replacement, then*

$$P\left(\lim_{n\to\infty} \frac{X_1 + X_2 + \ ... \ + X_n}{n} = \mu\right) = 1$$

One of the more entertaining consequences of the strong law of large numbers is the classic result below.

Example 6.1 *(Infinite Monkey Theorem) If an immortal monkey at a typewriter presses keys at random for eternity, show that the monkey will eventually produce the complete works of Shakespeare.*

Consider the infinite sequence of keys the monkey types. Break that sequence up into infinitely many disjoint blocks of length N, where N is the number of key presses required to produce the complete works of Shakespeare. Let A_i be the event 'the i^{th}

block of length N is an exact reproduction of the complete works of Shakespeare', and let X_i be a random variable defined by

$$X_i = \begin{cases} 1 & \text{if } A_i \text{ occurs} \\ 0 & \text{if } A_i \text{ doesn't occur} \end{cases}$$

Assuming the monkey selects each key uniformly at random, without any regard to previous choices, we have $P(X_i = 1) = (\frac{1}{k})^N$ and $P(X_i = 0) = 1 - (\frac{1}{k})^N$, where k is the number of keys on the typewriter. Furthermore, the events $A_1, A_2, \ldots, A_n$ form an independent sequence since the blocks of length N do not overlap.

Therefore, the random variables X_i are independent and identically distributed. In other words, they form a SRS of n values from a random variable X which takes value one with probability $(\frac{1}{k})^N$ and zero with probability $1 - (\frac{1}{k})^N$.

The expected value of X is $\mu = E(X) = 1 \cdot (\frac{1}{k})^N + 0 \cdot (1 - (\frac{1}{k})^N) = (\frac{1}{k})^N$. Thus, by the strong law of large numbers,

$$P\left(\lim_{n\to\infty} \frac{X_1 + X_2 + \ldots + X_n}{n} = \left(\frac{1}{k}\right)^N\right) = 1$$

Note that $\left(\frac{1}{k}\right)^N \neq 0$, so with probability one, $\lim_{n\to\infty} \frac{X_1+X_2+\ldots+X_n}{n}$ is not zero. This means that the sum $X_1 + X_2 + \ldots + X_n \to \infty$ as $n \to \infty$, since if this sum was bounded by some constant c, then $\frac{X_1+X_2+\ldots+X_n}{n} \leq \frac{c}{n} \to 0$. Therefore, we have $X_i = 1$ for infinitely many i, so with probability one, not only will the monkey produce the complete works of Shakespeare, it will do so infinitely many times!

The law of large numbers is frequently misunderstood. Consider a scenario in which a fair coin is repeatedly flipped. The law of large numbers states that as the number of flips grows larger and larger, the proportion of heads will converge to $\frac{1}{2}$ with probability one.

If the first five flips all result in tails, it might seem reasonable to use the law of large numbers to argue that the coin is 'due for heads', and so heads must now be more likely than tails, since the proportion of heads must converge to $\frac{1}{2}$. This is known as the gambler's fallacy.

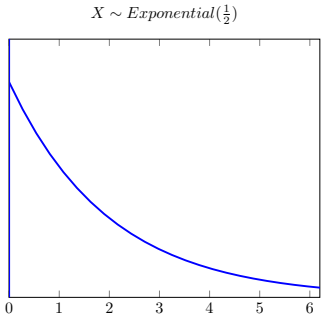
The key point is that as the number of flips grows very large, the influence of the first five flips on the proportion of heads becomes less and less significant. Suppose that after the first five flips, all of which are tails, heads and tails simply alternate, beginning with heads. After ten flips, the proportion of heads is $\frac{3}{10} = 30\%$, after one hundred flips, it's $\frac{48}{100} = 48\%$, and after one thousand flips, it's $\frac{498}{1000} = 49.8\%$. The coin does not need to become biased towards heads to bring the proportion of heads closer to $\frac{1}{2}$. After many flips, the initial streak of five tails simply becomes less and less significant.

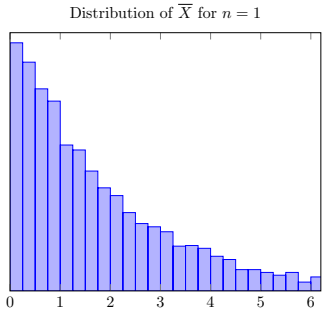
6.2 The Central Limit Theorem

In the last section, two extremely important general results were established. We found expressions for the mean and standard deviation of the sampling distribution of $\overline{X}$ (given in Corollary 6.1) which hold for any random variable X whose mean μ and standard deviation σ are defined, and we showed the sample mean will converge to the population mean as the sample size grows without bound.

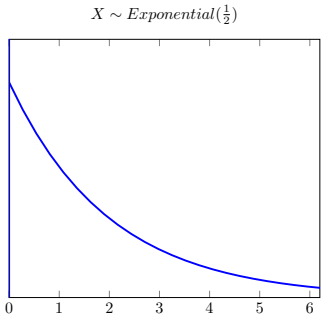
In other words, the sampling distribution of $\overline{X}$ will cluster tighter around the population mean μ as the sample size grows larger and samples which have sample means far from μ become more and more rare.

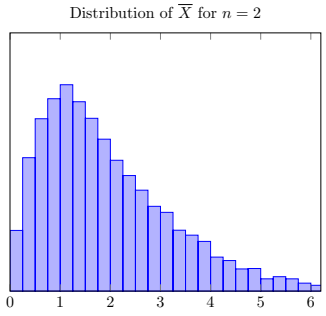
What about the shape of the sampling distribution of $\overline{X}$? What happens as the sample size increases? Consider the population distribution $X \sim Exponential(\frac{1}{2})$. Let's approximate the sampling distribution of $\overline{X}$ by taking ten thousand SRSs and calculating their means. With a sample size of one, the result should look just like the population distribution, since the mean of a single observation is just that observation itself.



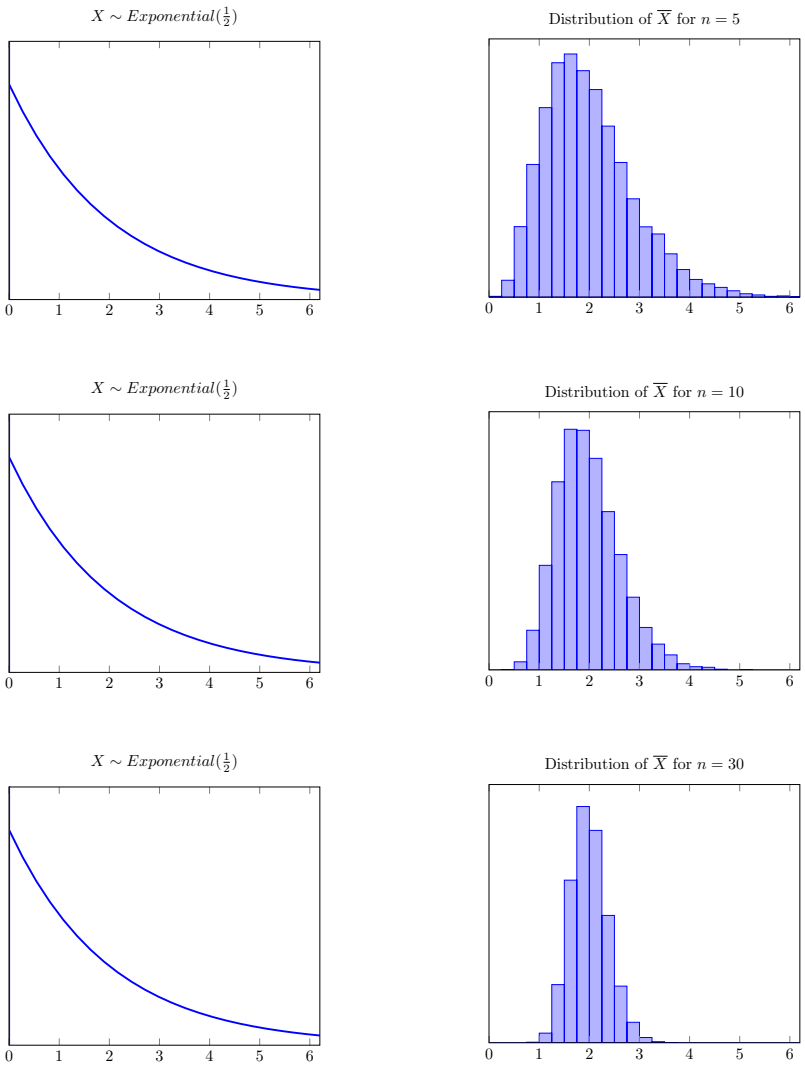


What happens when we increase the sample size to two? Let's take ten thousand samples, each consisting of two independent observations of $X \sim Exponential(\frac{1}{2})$, and compute the mean of each sample. The resulting distribution is given below.





If we continue to increase the sample size, taking ten thousand random samples of larger and larger sizes from the same population distribution, and computing the mean of each, how will the distribution of these means change?



Each of the sampling distributions has mean $\mu_{\overline{X}} = 2$, the same as the population distribution, and the standard deviation becomes smaller and smaller as n increases, as Corollary 6.1 assures us it will (since $\sigma_{\overline{X}} = \sigma/\sqrt{n} \to 0$ as $n \to \infty$).

Moreover, there is a striking trend in the shapes of the sampling distributions. As the sample size increases, the distribution looks less like an exponential distribution, and more like a normal distribution (albeit a narrow one).

It's not a coincidence that this occurred in our particular example. In fact, this occurs for almost *any* population! This is a hugely important result, and it's why the normal distribution plays such a central role in statistics. It implies that if some quantity is an average of a random sample, and the sample size was large enough, then we know that the sampling distribution it was drawn from must be approximately normal *regardless of the population the sample was drawn from!*

Theorem 6.3 *(Central Limit Theorem) Suppose X is a random variable with mean μ and standard deviation σ, and let $\overline{X}$ denote the mean of a sample of n values drawn from the distribution of X. Then the sampling distribution of $\overline{X}$ approaches $Normal(\mu_{\overline{X}}, \sigma_{\overline{X}})$ as $n \to \infty$.*

Example 6.2 *Suppose twenty numbers in the interval $[0, 1]$ are chosen uniformly at random. What is the probability their sum is higher than eight?*

In order to use the central limit theorem, we need to formulate this question in terms of the mean of a sample. In this case, we have a sample of $n = 20$ values drawn from $X \sim Uniform([0, 1])$. Observe that

$$X_1 + X_2 + ... + X_{20} > 8 \quad \to \quad \frac{X_1 + X_2 + ... + X_{20}}{20} > \frac{8}{20} = \frac{2}{5}.$$

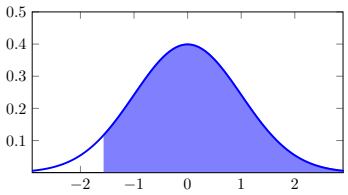
Thus, to answer this question, we need to find the probability of obtaining a sample of twenty values with a mean $\overline{X}$ higher than $\frac{2}{5}$. Using Corollary 6.1 and Proposition 4.6 The mean and standard deviation of $\overline{X}$ are given by

$$\mu_{\overline{X}} = \mu = \frac{1}{2} \quad \text{and} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{1}{12}(1-0)^2}}{\sqrt{20}} = \frac{1}{\sqrt{240}} \simeq 0.0645$$

The central limit theorem states that the distribution of $\overline{X}$ can be approximated by $Normal(0.5, 0.0645)$, so the probability of obtaining a sample with $\overline{X}$ higher than $\frac{2}{5} = 0.4$ can be calculated as follows.

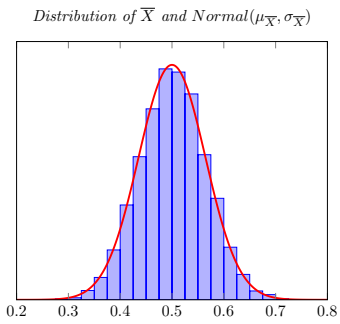
$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{0.4 - 0.5}{0.0645} = \frac{-0.1}{0.0645} \simeq -1.55$$

$$\begin{aligned} P(\overline{X} > \tfrac{2}{5}) &= P(Z > -1.55) \\ &= 1 - P(Z \leq 1.55) \\ &= 1 - 0.0603 \\ &= 0.9397 \simeq 94\% \end{aligned}$$



Note that the central limit theorem does not tell us how closely the distribution of $\overline{X}$ is approximated by $Normal(0.5, 0.0645)$. Let's take ten thousand samples and plot the resulting sampling distribution for $\overline{X}$, along with the normal distribution given by the central limit theorem.

If the two distributions match well, then we'll know that our answer above is fairly accurate. If the sampling distribution is still far from being normal, then we'll have to disregard our answer.



The distribution of $\overline{X}$ for samples of size twenty drawn from $X \sim Uniform([0,1])$ is extremely well approximated by a normal distribution, so we can be sure our answer is quite accurate.

How Close is Close Enough?

The central limit theorem states that the sampling distribution of $\overline{X}$ will converge to a normal distribution as the sample size $n \to \infty$, but how fast does this convergence occur? How large must the sample size be so we can simply use a normal distribution for calculations and have accurate results?

The answer varies with the population distribution. Symmetric population distributions with thin tails (tails which decay exponentially) usually yield sampling distributions for $\overline{X}$ which are very close to normal even with a very small sample size, while asymmetric population distributions with fat tails (tails which decay polynomially) can give sampling distributions for $\overline{X}$ which don't look normal until the sample size becomes extremely large.

As a rule of thumb, we'll use the guidelines below to determine if the sampling distribution of $\overline{X}$ is sufficiently close to normal that we can justify approximating it by a normal distribution.

(i) If the population distribution is normal, any sample size is fine.

(ii) If the population distribution is roughly symmetric and not fat-tailed, $n \geq 30$.

(iii) If the population distribution is significantly skewed or fat-tailed, $n \geq 50$.

There is nothing mathematically important about these values, or similar values you'll find in other textbooks. We're simply establishing a clear baseline.

In actual practice, to ensure the sampling distribution of $\overline{X}$ is sufficiently close to a normal distribution, statisticians either rigorously derive error bounds from some assumptions on the form of the population distribution, or use statistical software to take large numbers of samples from an educated guess at the population distribution and approximate the sampling distribution empirically, as was done to generate the graphics at the beginning of this section.

Example 6.3 *If a biased coin with a 55% probability of heads is flipped thirty times, what is the probability that the number of heads observed in this sample is less than the number of tails?*

If we call a head a success, then we have a sample of 30 values $X_1, X_2 \ldots, X_{30}$ *from* $X \sim Bernoulli(0.55)$. *If more tails than heads were observed, this means that more than half of the* X_i *are zeros, so* $\overline{X} = \frac{X_1+X_2+\ldots+X_{30}}{30} < \frac{15}{30} = 0.5$.

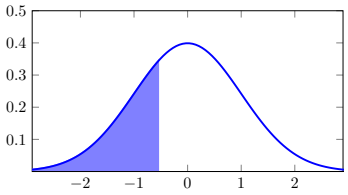
Note $X \sim Bernoulli(0.55)$ *is roughly symmetric, so by the central limit theorem, the distribution of* $\overline{X}$ *is approximately normal for samples of* $n = 30$ *observations. The mean and standard deviation of the sampling distribution are given by*

$$\mu_{\overline{X}} = \mu = 0.55 \quad \textit{and} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{\sqrt{0.2475}}{\sqrt{30}} \simeq 0.091$$

Now we can calculate the probability of obtaining a sample mean $\overline{X} < 0.5$.

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{0.5 - 0.55}{0.091} = \frac{-0.05}{0.091} \simeq -0.549$$

$$\begin{aligned} P(\overline{X} < 0.5) &= P(Z < -0.549) \\ &= 0.2905 \simeq 29\% \end{aligned}$$



Therefore, there is a 29% *chance out sample will have more tails than heads. What do you expect will happen if we take a larger sample, say* $n = 100$ *flips? Try redoing the problem in that case.*

Note that in the example above, if we let Y be the number of heads in 30 flips, then $Y \sim Binomial(30, 0.55)$, so we can obtain a more accurate answer (in a more elementary but more tedious way) by evaluating

$$P(0 \leq Y \leq 14) = \sum_{k=0}^{14} P(Y = k) = \sum_{k=0}^{14} \binom{30}{k} (0.55)^k (0.45)^{30-k} \simeq 0.23.$$

So our answer above is off by about 6%. The vast majority of this error is not in fact due to the use of a normal distribution as an approximation, but simply to the translation from discrete to continuous values of $\overline{X}$. We used the central limit theorem to calculate the probability $\overline{X} < 0.5 = \frac{15}{30}$, but values between $\overline{X} = \frac{14}{30}$ and $\overline{X} = \frac{15}{30}$ never occur since thirty flips always result in a whole number of heads. In the former case, there are more tails than heads, and in the latter case there aren't.

If we split the difference and use $\overline{X} = \frac{14.5}{30}$ in our calculations with the central limit theorem, we will in fact arrive at the same answer we obtained just above, 23%. This modification is often referred to as a *continuity correction*.

6.3 Confidence Interval for a Mean

Now let's put the Central Limit Theorem to work. We're interested in knowing the average age of students at Champlain College. To estimate this quantity, we take a random sample of 50 students and record their ages. It turns out that the mean age of students in the sample is 18.3 years.

The sample mean, $\overline{x} = 18.3$, is a value of the random variable $\overline{X}$. Since the sample is a random sample of a sufficiently large size, the Central Limit Theorem tells us that the distribution of $\overline{X}$ is $Normal(\mu_{\overline{X}}, \sigma_{\overline{X}})$. Standardizing yields

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{18.3 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}.$$

Now suppose we happen to know that the standard deviation for ages of students at Champlain College is 2.7 years. The mean and standard deviation of $\overline{X}$ are (by Corollary 6.1) $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 2.7/\sqrt{50} \simeq 0.381$.

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{18.3 - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{18.3 - \mu}{0.381}$$

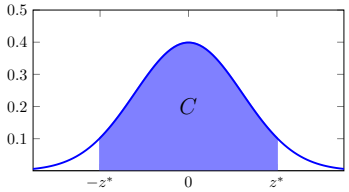
We've seen that about 95% of all values of Z fall into the interval $(-2, 2)$, and hence we should be about 95% confident that

$$\begin{gathered} -2 < \tfrac{18.3-\mu}{0.381} < 2 \\ -0.762 < 18.3 - \mu < 0.762 \\ -19.062 < -\mu < -17.538 \\ 17.538 < \mu < 19.062. \end{gathered}$$

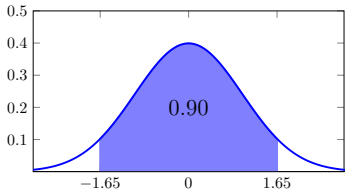
This range of values is called a confidence interval for the parameter μ, which in this case represents the mean age of all students at Champlain College. We can't determine the exact value of μ without collecting information from every single student. However, from a random sample of only fifty students, we can assert that $17.5 < \mu < 19.1$ with 95% confidence.

Confidence intervals are so useful and pervasive in applications of statistics, it's worth our while to set up some terminology to make constructing them as fast and simple as possible.

Definition 6.1 *Given a confidence level $C \in (0, 1)$, the critical value corresponding to that confidence level is the value z^* with $P(-z^* < Z < z^*) = C$.*



Example 6.4 *The critical value for a confidence level of* 90% *is* $z^* = 1.65$*, since* $P(-1.65 < Z < 1.65) = P(Z < 1.65) - P(Z < -1.65) = 0.9505 - 0.495 = 0.9$*.*



When using the Z-table, we can find z^* *by noting that if* $P(-z^* < Z < z^*) = 0.9$*, then by symmetry,* $P(Z < z^*) = 0.95$*. Thus, we're looking for the value of* z *that has an area of 0.95 to its left.*

If we follow through the calculation we made in the introduction to this section with a general critical value z^* and a general sample mean $\overline{X}$, we'll arrive at a formula for a confidence interval for the mean μ of any population from which we might draw a sample. With a probability of C, we have

$$\begin{aligned}
-z^* &< Z < z^* \\
-z^* &< \tfrac{\overline{X}-\mu_{\overline{X}}}{\sigma_{\overline{X}}} < z^* \\
-z^*\sigma_{\overline{X}} &< \overline{X} - \mu < z^*\sigma_{\overline{X}} \\
-\overline{X} - z^*\sigma_{\overline{X}} &< -\mu < -\overline{X} + z^*\sigma_{\overline{X}} \\
\overline{X} - z^*\sigma_{\overline{X}} &< \mu < \overline{X} + z^*\sigma_{\overline{X}} \\
\overline{X} - z^*\tfrac{\sigma}{\sqrt{n}} &< \mu < \overline{X} + z^*\tfrac{\sigma}{\sqrt{n}}.
\end{aligned}$$

Note that this calculation only makes sense when $\overline{X}$ is normally distributed, but the Central Limit Theorem tells us this is indeed the case as long as our sample is a SRS of a sufficiently large size. The quantity we add to and subtract from $\overline{X}$ in the last line is called the margin of error.

Definition 6.2 *Given a SRS of* n *values drawn from any distribution and a confidence level* $C \in (0, 1)$*, the margin of error for* $\overline{X}$ *at confidence level* C *is*

$$\boxed{E = z^*\frac{\sigma}{\sqrt{n}}}$$

This value represents the largest possible gap between a sample mean $\overline{X}$ and the population mean μ that is observed at the confidence level C. For example, if $E = 2.2$ when $C = 0.9$, this means there is a 90% probability that a SRS has a mean which differs from the true population mean by less than 2.2.

With our definitions established, we can describe the procedure for constructing a confidence interval for a population mean μ based on any SRS of n values with sample mean $\overline{X}$ in three steps.

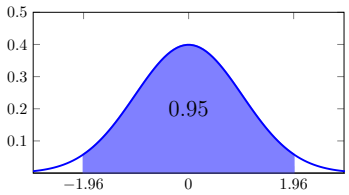
(i) Find the critical value z^* for the desired confidence level C.

(ii) Compute the margin of error E.

(iii) The confidence interval is given by $\overline{X} - E < \mu < \overline{X} + E$.

Example 6.5 *The standard deviation for heights of all men in Canada is known to be around 2.9 inches. In a random sample of 37 Canadian men, the average height was 5'9". Use this information to construct a 95% confidence interval for the average height of all Canadian men.*

The critical value z^* is defined by $P(-z^* < Z < z^*) = 0.95$, so the area on the left of z^* must be $P(Z < z^*) = 0.975$, and consulting the Z-table yields $z^* = 1.96$.



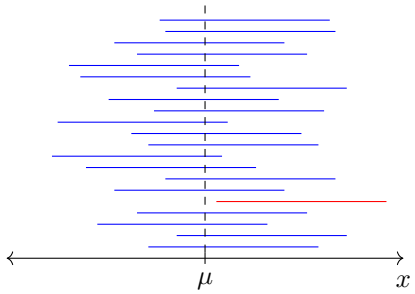
The margin of error is then given by $E = z^* \frac{\sigma}{\sqrt{n}} = 1.96 \frac{2.9}{\sqrt{37}} \simeq 0.934$. In inches, the sample mean 5′9" is 69", so we obtain the interval

$$69 - 0.934 < \mu < 69 + 0.934 \quad \rightarrow \quad 68.066 < \mu < 69.934$$

Thus, with 95% confidence, the mean height of all Canadian men must be between 68.066" (about 5'8") and and 69.934" (about 5'10").

Note that when we say 'with 95% confidence, $68.066 < \mu < 69.934$', this means there a 95% chance that we have selected a sample which will produce a confidence interval containing the true mean height of the population, μ.

The mean height of all Canadian men is some fixed single value, we just don't know what it is. Thus, the probability it's inside our particular confidence interval is either 100% or 0% (it is or it isn't). But if we were to take many random samples and compute a 95% confidence interval from each, we would expect that, typically, nineteen of every twenty samples would produce intervals that contain the true value of μ (since 19/20 is 95%).



Notice that as the confidence level becomes larger, the corresponding critical value z^* grows to enclose that larger area, which increases the margin of error, E. On the other hand, the value of E is a decreasing function of the sample size, n, so a larger sample will lead to a smaller confidence interval, giving a more accurate estimate of the parameter μ.

Proposition 6.1 *To reduce the margin of error, E and produce tighter bounds on μ, we can either decrease the confidence level, C, or increase the sample size, n.*

6.4 Confidence Interval for a Proportion

What if we're interested in estimating the proportion of individuals in a population that have some property of interest? For example, a pollster might want to estimate the proportion of all voters in some area that intend to vote for a particular political party. We can produce a confidence interval for such a proportion from a random sample using the ideas we developed in the last section.

Each individual in the population either has the desired property or does not, so if the proportion of individuals in the population with the desired property is $p \in [0, 1]$ (and hence the proportion without is $1 - p$), the population distribution we're interested in is $X \sim Bernoulli(p)$ for some unknown parameter p. We'll call p the population proportion. The expected value of its distribution is $E(X) = p$, hence the population mean is also $\mu = p$, as we saw in Section 3.5.

Notice that if we sample from a Bernoulli distribution, then the sample mean $\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$ is actually the proportion of individuals in our sample with the desired property (since X_i was drawn from a Bernoulli distribution, so it take the value 1 when the i^{th} individual has the property, and 0 if not). Thus, we're counting the number of individuals in our sample with the property and dividing by the sample size, so the sample mean $\overline{X}$ is in fact equal to the sample proportion in this context. To remind ourselves it's a proportion we can write $\widehat{p} = \overline{X}$.

In summary, if our population distribution is $X \sim Bernoulli(p)$, then the population mean and sample mean are equal to the population proportion and sample proportion respectively, that is, $p = \mu$, and $\widehat{p} = \overline{X}$. This means we can construct confidence intervals for a proportion in exactly the same way we made confidence intervals for the mean in the last section.

In Section 3.5, we saw that the variance of a Bernoulli distribution is given by the formula $\sigma^2 = p(1 - p)$, so the margin of error E becomes

$$E = z^* \frac{\sigma}{\sqrt{n}} = z^* \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

and although we can't know the value of $\sqrt{p(1-p)}$ (because p is the unknown we're trying to estimate), we can bound it.

Proposition 6.2 *If $p \in [0,1]$, then $0 \leq \sqrt{p(1-p)} \leq \frac{1}{2}$, and therefore the margin of error for a sample proportion $\widehat{p}$ satisfies*

$$\boxed{E \leq z^* \frac{1}{2\sqrt{n}}}$$

Pf. Define $f(p) = p(1-p)$. Notice that to max/minimize $\sqrt{p(1-p)}$ on $[0,1]$, we can max/minimize $f(p)$ on $[0,1]$. Since f is continuous and differentiable on $[0,1]$, it will attain its max/minimum values at a critical point or at an endpoint.

$$f'(p) = (1)(1-p) + (p)(-1) = 1 - 2p$$

Therefore, the only critical point of f is $p = \frac{1}{2}$. Evaluating f at this point and the two endpoints of $[0,1]$ yields

p	$f(p)$
0	0
$\frac{1}{2}$	$\frac{1}{4}$
1	0

Therefore, f attains its maximum value at $p = \frac{1}{2}$, and hence

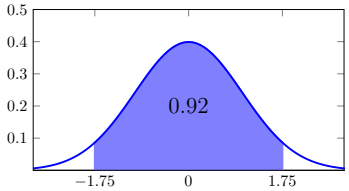
$$E = z^* \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq z^* \frac{\sqrt{\frac{1}{2}(1-\frac{1}{2})}}{\sqrt{n}} = z^* \frac{1}{2\sqrt{n}}$$

□

We can use this result to construct confidence intervals for a population proportion, but keep in mind we have an upper bound, not the true value of E. When we construct a 95% confidence interval, there should be a 95% probability of selecting a sample that results in an interval containing the parameter we're estimating. If we use the bound on E that was just derived, this probability may in fact be larger than 95%, that is, our interval will generally be be wider than is necessary.

Example 6.6 *Suppose that a pollster is interested in knowing what proportion of voters in a certain riding plan to vote NDP. In a random sample of 817 voters from this riding, 227 plan to vote NDP. Construct a 92% confidence interval for the proportion of all voters in this riding who plan to vote NDP.*

The critical value z^* is defined by $P(-z^* < Z < z^*) = 0.92$, so the area on the left of z^* must be $P(Z < z^*) = 0.96$, and consulting the Z-table yields $z^* = 1.75$.



The margin of error satisfies $E \leq z^* \frac{1}{2\sqrt{n}} = 1.75 \frac{1}{2\sqrt{817}} \simeq 0.031$*, and since the sample proportion is* $\widehat{p} = \frac{227}{817} = 0.278$*, we have*

$$0.278 - 0.031 < p < 0.278 + 0.031 \quad \rightarrow \quad 0.247 < p < 0.309.$$

Thus, with 95% *confidence, the proportion of voters in the riding who plan to vote NDP is between* 24.7% *and* 30.9%*.*

Note that if you look up the margin of error formula for estimating a proportion in other sources you'll likely encounter the formula

$$E = z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

This is the same error bound we derived at the beginning of this section, but the unknown population proportion p has been replaced by the sample proportion $\widehat{p}$. As the sample size increases, we know $\widehat{p} \to p$ by the law of large numbers, but for any fixed n, they will differ, so to properly justify this error bound requires more work, as this difference introduces a new source of uncertainty we have not quantified.

6.5 Confidence Interval for a Mean (σ Unknown)

Our procedure for constructing a confidence interval for a population mean μ requires knowledge of the population standard deviation σ. This is to be expected, as means of samples from a population with a large standard deviation will be more variable more than means of samples from a population with a low standard deviation. Unfortunately, this requirement is problematic in practice since σ is a parameter of the population distribution, so is typically unknown.

We got around this requirement when estimating a proportion by establishing an upper bound on the population standard deviation $\sigma = \sqrt{p(1-p)}$, but no such upper bound exists in general when we're estimating a mean.

If we can't find an upper bound the value of σ, can we estimate it in some other way? Well, the best estimate of the population we have access to is our sample, so what if we simply use the standard deviation of our sample in place of the standard deviation of the population?

Definition 6.3 *The (Bessel corrected) sample variance is* $S^2 = \frac{1}{n-1} \sum_i (X_i - \overline{X})^2$ *and the sample standard deviation,* S*, is the square root of the sample variance.*

Example 6.7 *Find the mean and standard deviation of the sample of five values* $X_1 = 2$*,* $X_2 = 0$*,* $X_3 = 7$*,* $X_4 = 3$*,* $X_5 = 3$*, drawn from an unknown distribution.*

$$\overline{X} = \frac{1}{n} \sum_i X_i = \frac{1}{5}(2 + 0 + 7 + 3 + 3) = 3$$

$$S = \sqrt{\frac{1}{n-1} \sum_i (X_i - \overline{X})^2} = \sqrt{\frac{1}{4}((-1)^2 + (-3)^2 + (4)^2 + (0)^2 + (0)^2)} = \sqrt{\frac{13}{2}}$$

Note that when calculating S^2 and S, we don't divide by the sample size, n, as in Definition 3.6, but instead by $n-1$. This change is known as Bessel's correction, and it's done to obtain an unbiased estimator of the population variance. The formula in Definition 3.6 systematically underestimates the population variance.

Theorem 6.4 *The Bessel corrected sample variance* $S^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$*, is an unbiased estimator of the population variance, that is,* $E(S^2) = \sigma^2$*.*

Pf. This follows from a long and careful calculation using properties of expected value. One key fact to recall is that $Var(X) = E[X^2] - (E[X])^2$ for any random variable X, and hence $E[X^2] = Var(X) + (E[X])^2$.

$$\begin{aligned}
E(S^2) &= E\left(\tfrac{1}{n-1}\sum_i (X_i - \overline{X})^2\right) \\
&= \tfrac{1}{n-1} E\left(\sum_i (X_i - \overline{X})^2\right) \\
&= \tfrac{1}{n-1} E\left(\sum_i \left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right)\right) \\
&= \tfrac{1}{n-1} E\left(\sum_i X_i^2 - \sum_i 2X_i\overline{X} + \sum_i \overline{X}^2\right) \\
&= \tfrac{1}{n-1} E\left(\sum_i X_i^2 - 2\overline{X}\sum_i X_i + n\overline{X}^2\right) \\
&= \tfrac{1}{n-1} E\left(\sum_i X_i^2 - 2\overline{X}\cdot n\overline{X} + n\overline{X}^2\right) \\
&= \tfrac{1}{n-1} E\left(\sum_i X_i^2 - n\overline{X}^2\right) \\
&= \tfrac{1}{n-1} \left(\sum_i E[X_i^2] - E[n\overline{X}^2]\right) \\
&= \tfrac{1}{n-1} \left(nE[X^2] - nE[\overline{X}^2]\right) \\
&= \tfrac{1}{n-1} \left(n(Var(X) + (E[X])^2) - n(Var(\overline{X}) + (E[\overline{X}])^2)\right) \\
&= \tfrac{1}{n-1} \left(n(\sigma^2 + \mu^2) - n(\tfrac{\sigma^2}{n} + \mu^2)\right) \\
&= \tfrac{1}{n-1} \left(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right) \\
&= \tfrac{1}{n-1} \left((n-1)\sigma^2\right) = \sigma^2
\end{aligned}$$

□

In practice, Bessel's correction is used almost universally. Try calculating the standard deviation of a small list of values using a spreadsheet on a computer. Is the calculation done with or without Bessel's correction?

If we use the sample standard deviation, s, to estimate the standard deviation in our population, σ, then when we standardize a sample mean $\overline{X}$, instead of the standardized value

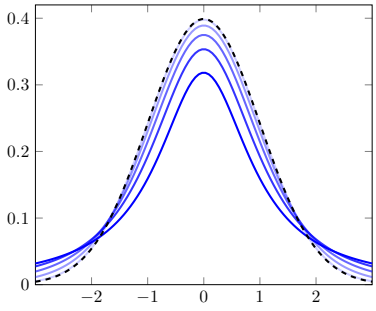
$$\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}, \quad \text{we'll have the estimate} \quad \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \simeq \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}.$$

Our construction of a confidence interval for the population mean μ was justified by the Central Limit Theorem, which tells us the distribution of the quantity on the left approaches $Z \sim Normal(0, 1)$ as the sample size n grows (and is exactly equal to $Z \sim Normal(0, 1)$ when the population distribution is normal). But what about our estimate on the right? If we repeatedly draw samples and calculate this value, what kind of distribution will we obtain?

Definition 6.4 *Suppose we draw a SRS $X_1, X_2, \ldots, X_n$ from a normally distributed population with mean μ, and define the T-score of the sample as*

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

where S is the sample standard deviation, as in Definition 6.3. Then T is a continuous random variable whose distribution is called Student's T-distribution (or simply the T-distribution) with $n - 1$ degrees of freedom.



Above in blue are plots of probability density functions for T-distributions with one, two, four, ten, and one hundred degrees of freedom, shaded lighter as the degree of freedom grows. The standard normal distribution is shown in dashed black.

The T-distribution has fatter tails and less area near the center than the standard normal distribution. This additional variability appears because the sample standard deviation S differs from one sample to the next, which adds a source of variability to T-scores which is not present in Z-scores. As the sample size grows, the sample standard deviation S becomes a better estimator of the population standard deviation σ, and the T-distribution slowly approaches the standard normal distribution. Notice that once the degree of freedom reaches one hundred, the T-distribution and the standard normal distribution are effectively identical.

If we can find areas under the T-distribution, then the same procedure we've been using to construct confidence intervals with the population standard deviation can be modified to work with the sample standard deviation.

Proposition 6.3 *Let T be a random variable whose distribution is the T distribution with $n-1$ degrees of freedom, and take t^* so that $P(-t^* < T < t^*) = C$.*

If we define $E = t^* \frac{S}{\sqrt{n}}$, then $P(\mu \in (\overline{X} - E, \overline{X} + E)) = C$, where $\overline{X}$ is the mean of a SRS drawn from a normal distribution with mean μ.

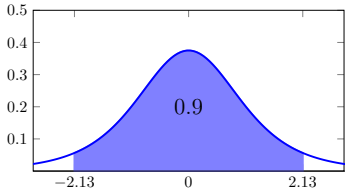
Pf. By definition of t^*, with probability C we have

$$\begin{aligned} -t^* < T < t^* \\ -t^* < \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} < t^* \\ -t^* \frac{S}{\sqrt{n}} < \overline{X} - \mu < t^* \frac{S}{\sqrt{n}} \\ -\overline{X} - t^* \frac{S}{\sqrt{n}} < -\mu < -\overline{X} + t^* \frac{S}{\sqrt{n}} \\ \overline{X} - t^* \frac{S}{\sqrt{n}} < \mu < \overline{X} + t^* \frac{S}{\sqrt{n}}. \end{aligned}$$

□

Example 6.8 *A sample of five values is drawn from a normal distribution with unknown mean and standard deviation. These values are 43, 27, 30, 22, and 38. Construct a 90% confidence interval for the mean of the normal distribution.*

Computing the mean and (Bessel corrected) standard deviation of the sample yields $\overline{X} = 32$ and $S = 8.456$. We can now construct a confidence interval for the population mean μ just as we've done before in Section 6.3, but using s in place of σ and the T-distribution in place of the standard normal distribution.



We compute $t^* = 2.13$ using a table of values for the t-distribution (with 4 degrees of freedom), then $E = t^* \frac{s}{\sqrt{n}} = 2.13 \cdot \frac{8.456}{5} = 3.60$, and therefore with 90% confidence we have

$$32 - 3.60 < \mu < 32 + 3.60 \quad \rightarrow \quad 28.40 < \mu < 35.60$$

Robustness

Notice that the T-distribution is defined as the distribution of T-scores of random samples drawn from a normal distribution. In practice, however, we would rarely

know enough about the population we are studying to be able to determine whether the distribution of the random variable we are interested in is normal.

Fortunately, whenever the hypotheses of the central limit theorem hold, the distribution of T-scores of random samples drawn from any distribution will, in fact, eventually converge to the T-distribution [5].

Even with smaller sample sizes, statisticians like to say that the procedure illustrated above in Example 6.8 is 'robust to non-normality'. This means that the assumption that we're sampling from a normal distribution is necessary for strict correctness, but that breaking this assumption will typically introduce only a small source of error. We also know this error can be completely eliminated by having a sufficiently large sample size. With small samples though, one can be justifiably suspicious, especially if there's reason to believe the distribution of the population being studied is fat-tailed or very asymmetric.

Example 6.9 *In a study of a particular species of maple in a provincial park, the height of every tree in a random sample of 38 mature trees was measured. If the sample mean and sample standard deviation for the heights of the trees were* 21.3 *metres and* 4.9 *metres respectively, construct a* 90% *confidence interval for the mean height of all trees of this species in the park.*

We have $\overline{X} = 21.3$*,* $S = 4.9$*, and* $n = 38$*. Using the T-table, we can see that in a T-distribution with 37 degrees of freedom,* $P(-1.69 < T < 1.69) = 0.9$*, so the critical value is* $t^* = 1.69$*.*

We can then compute the margin of error $E = t^* \frac{s}{\sqrt{n}} = 1.69 \frac{4.9}{\sqrt{38}} = 1.343$*. We can then write the confidence interval*

$$21.3 - 1.343 < \mu < 21.3 + 1.343 \quad \rightarrow \quad 19.957 < \mu < 22.643.$$

Therefore, with 90% *confidence, the mean height of all mature trees of this species in the park is between* 19.9 *and* 22.7 *metres.*

In examples like the one above, where the distribution our sample is drawn from is somewhat bell-shaped, without fat tails (think about what kind of shape the distribution of heights of mature trees might have), confidence intervals constructed using the T-distribution are typically very accurate for all but extremely small sample sizes.

Bibliography

[1] D. Bertsekas and J. Tsitsiklis, *Introduction to Probability (2nd Ed)*, Athena Scientific, 2008.

[2] W. Feller, *An Introduction to Probability, Vol I (2nd Ed)*, Wiley, 1950.

[3] M. Gardner, *The Second Scientific American Book of Mathematical Puzzles and Diversions*, Simon & Schuster, 1954.

[4] S. Ghahramani, *Fundamentals of Probability (2nd Ed)*, Pearson, 2000.

[5] E. Giné, F. Götze, and D. Mason, *When is the student t-statistic asymptotically standard normal?*, The Annals of Probability, 25 (1997), pp. 1514–1531.

[6] A. Kolmogorov, *Foundations of the Theory of Probability (2nd English Ed)*, Chelsea Publishing Company, 1956.

[7] S. Selvin, *A problem in probability (letter to the editor)*, American Statistician, 29 (1975), p. 67.

[8] M. vos Savant, *Ask Marilyn*, Parade Magazine, 16 (1990).

Index